

智能之光：人机协作的经济管理研究新时代

胡诗云 易君健

摘要：以ChatGPT 和DeepSeek 为代表的人工智能大语言模型（简称大模型），正在对知识工作者的生产方式产生革命性的影响。本文面向经济管理学科的研究者，介绍大模型的技术原理、应用方式以及在科学研究全流程中的应用。本文首先从社会科学和大语言模型的本质出发，分析了认知自动化的边界，指出围绕理论工作的能力是人类科学家在人工智能时代的核心能力。随后，本文介绍大模型的基本原理、关键技术和发展趋势，经济管理研究者可以建立对人工智能技术的基本认知。接下来，本文介绍了通过API 访问、本地部署等大模型的具体应用方式。随后本文提出了大模型在经济管理研究全流程中的四种角色：参谋、助研、智能体和朋友。作为参谋，大模型充当研究者的思想伙伴，协助资料收集、深化文献理解、澄清概念并提供研究反馈；作为助研，大模型承担研究助理职能，处理文献整理、参考文献格式调整、梳理建模和推导等重复性任务；作为智能体，大模型本身成为研究对象与实验工具，能模拟人类决策行为、预测反应，并通过多智能体系统模拟社会互动；作为朋友，大模型超越学术角色，提供全方位非学术支持，包括职业发展建议、心理辅导和人际交往建议。大模型全面融入学术工作流程，不仅能通过自动化重复劳动提高研究效率，更能通过人机合作扩展人类思维的广度和深度，经济管理研究即将走向人机协作的新时代。

关键词：人工智能；经济学方法论；人机协同；大语言模型

JEL Codes: A11; B41; C45; D83

智能之光：人机协作的经济管理研究新时代

胡诗云，易君健*

2025 年 8 月

摘要：以 ChatGPT 和 DeepSeek 为代表的人工智能大语言模型（简称大模型），正在对知识工作者的生产方式产生革命性的影响。本文面向经济管理学科的研究者，介绍大模型的技术原理、应用方式以及在科学研究全流程中的应用。本文首先从社会科学和大语言模型的本质出发，分析了认知自动化的边界，指出围绕理论工作的能力是人类科学家在人工智能时代的核心能力。随后，本文介绍大模型的基本原理、关键技术和发展趋势，经济管理研究者可以建立对人工智能技术的基本认知。接下来，本文介绍了通过 API 访问、本地部署等大模型的具体应用方式。随后本文提出了大模型在经济管理研究全流程中的四种角色：参谋、助研、智能体和朋友。作为参谋，大模型充当研究者的思想伙伴，协助资料收集、深化文献理解、澄清概念并提供研究反馈；作为助研，大模型承担研究助理职能，完成文献整理、参考文献格式调整、数理建模和推导等常规任务；作为智能体，大模型本身成为研究对象与实验工具，能模拟人类决策行为、预测反应，并通过多智能体系统模拟社会互动；作为朋友，大模型超越学术角色，提供全方位非学术支持，包括提供职业发展建议、心理辅导和人际交往建议。大模型全面融入学术工作流程，不仅能通过自动化来提高研究效率，更能通过人机合作扩展人类思维的广度和深度，经济管理研究即将走向人机协作的新时代。

关键词：人工智能；经济学方法论；人机协同；大语言模型

JEL Codes: A11; B41; C45; D83

*胡诗云，北京大学国家发展研究院，博士研究生，电子邮箱：hushiyun@pku.edu.cn。易君健，北京大学国家发展研究院，教授，电子邮箱：junjian@nsd.pku.edu.cn。作者感谢张皓辰、尹晔、李欣元、王若璧、江弘毅、周子焜、吴浩哲、林轶凡、颜玮、赵家琪、曹治、李星宇、王玥琴、韩旭、周清源、王悦、李奕天、汪姝君、王鹏飞参与讨论并贡献提示词案例，文责自负。

目录

1	引言	4
2	人工智能时代的社会科学家	5
2.1	人工智能可以做什么？	6
2.2	人工智能做不了什么？	7
2.3	社会科学家与 AI 在知识创新中的角色	9
3	祛魅 AI：大模型的基本原理	10
3.1	大语言模型的定义和历史	11
3.2	大语言模型的数学结构	12
3.3	大语言模型的训练	17
3.4	采样和推理	21
3.5	大模型下游应用	23
3.6	大语言模型的总体发展趋势	26
4	应用大模型的四种方式	27
4.1	通过网页访问	27
4.2	通过 API 访问	29
4.3	大模型本地部署	32
4.4	整合大模型的其它应用	33
5	作为参谋的 AI	34
5.1	澄清研究思路	35
5.2	信息资料的收集	36
5.3	文献的批判性阅读	36
5.4	研究反馈与改进	40
5.5	语言润色	41
6	作为助研的 AI	42
6.1	大模型作为助研时的一般技巧	42
6.2	完成重复性工作	43
6.3	数理建模与推导	44
6.4	撰写代码	45
6.5	提取事实信息	46

6.6	构造经济指标	47
7	作为智能体的 AI	50
7.1	单智能体模拟个体行为	50
7.2	多智能体模拟社会互动	51
7.3	智能体模拟的问题和局限性	51
8	作为朋友的 AI	52
8.1	提供职业发展建议	53
8.2	提供心理辅导	53
8.3	提供人际交往建议	54
9	结论	54
9.1	进一步阅读建议	54
9.2	技术特点与应用启示	55
9.3	将大模型融入研究全流程	56

1 引言

智能技术革命正重塑全球学术生态。人工智能大语言模型¹与传统自动化有本质区别：大语言模型实现了“认知工作的自动化”（Automation of cognitive work）（Korinek, 2023）。它们能够理解复杂概念、生成连贯文本并进行多步推理。如果说上一轮自动化取代了蓝领工人并催生了德鲁克所谓“知识工作者”，那么大语言模型就直接击中了知识工作者任务的核心。

经济学家作为知识工作者，毫无疑问处在这次变革的中心。经济学家的日常工作，包括数据分析、编程和专业写作等，恰恰都是大语言模型擅长而且仍在快速进步的领域。如果我们采用 O*NET 对于经济学家的任务描述，请 GPT-4o 来评判有多少会受到人工智能的影响，经济学家已经有 64% 的任务暴露于人工智能的影响之下。人工智能的飞速发展，一方面使我们好奇：究竟是怎样的技术，能够具备如此令人惊叹的文字输出和思考能力？它更促使我们思考：当认知辅助工具已能处理大量学术任务时，我们应如何重新定义经济学家的角色与价值？更重要的是，如何将这些工具整合到研究全流程中，使之成为智力放大器而非简单替代品？

幸运的是，我们有机会找到与人工智能合作的方式。最直接地，大模型可以帮助我们自动化许多“脏活”，比如修改格式、撰写文章、润色代码。除此之外，我们还可以把大模型作为自己可以随时与之对话的参谋，为我们提供智识上的启发。除了经济学家的日常工作之外，大语言模型也为经济学研究提供了新的分析工具。例如，多模态大模型为另类数据提供了全新的研究方法。这些技术能够处理传统经济学方法难以量化的信息，如图像、音频和非结构化文本，为经济学研究拓展了新的视野和分析维度。同时，大模型涌现出的行为能力本身，也为经济学提供了新的研究对象。这些人工智能系统展现出的决策模式、信息处理机制和策略行为，为经济学理论的创新应用与检验提供了前所未有的场景。

本文旨在向经济学家、管理学家和其它社会科学研究者全面地介绍大语言模型的技术和应用。本文首先从社会科学知识生产的一般过程出发，讨论大语言模型如何改变知识的生产函数。我们讨论了大语言模型在发现现象、提出问题、构建理论、检验假说再到传播与应用这些研究环节中的作用，并在每个环节中识别了人类的不可替代之处。我们发现，大模型可以在研究的各个环节实现常规认知任务的自动化执行，在收集数据、文献汇总、数理推导、数据分析、文字撰写等方面提高社会科学研究的执行效率，扩充社会科学家的知识库、认知边界和表达能力。但是，大语言模型作为语言统计模型的本质，决定了它没有提出理论问题的动机，也缺乏从现实数据中抽象出新概念并构建理论的能力。这部分讨论这为社会科学研究中应用大语言模型提供了思考框架，并且对更好的人机协作提出了启示。

随后本文介绍了大语言模型的技术原理。事实上，大语言模型并不神秘，它只不过是利用大量真实文本所训练出的“下一词预测器”（Next Token Predictor）。这一部分还介绍了与大模型相关的一系列术语。透过这些术语，我们可以更好地看出大模型日新月异发展背后的本质，从而更加客观地对待大模型取得的种种进展，破除对它们的迷信，并且更加正确、高效率地利用人工智能。

接下来，本文介绍应用大模型的四种方式：通过网页访问、通过 API 访问、本地部署以及使用整合了大模型的其它应用，包括代码编辑器以及大模型知识库。对于一般咨询的场景，使用网页访问即可满足要求；而

¹从概念的包含关系上看，人工智能（Artificial Intelligence, AI）包含了生成式人工智能，生成式人工智能又包含了大语言模型，简称大模型。本文主要讨论大语言模型的影响，因此若上下文中无特殊说明，本文中的“AI”、“大模型”均指代大语言模型。

对于需要大规模文本分析的应用，则需要通过 API 访问；如果有进一步的保密需求，则需要考虑本地部署。随着相关计算框架的成熟，本地部署大模型、微调大模型已经不是高科技公司的专利，而是每一个社会科学研究者都能运用的工具。

在介绍完这些技术基础之后，本文展示大模型在研究全过程中的应用。人工智能大模型在研究过程中将会扮演四种角色：参谋、助研、智能体，以及朋友。作为参谋，AI 可以协助创意生成、文献阅读、研究设计等工作，增强研究者的批判性思维；作为助研，AI 能够支持数理分析与数据处理；作为智能体，AI 可以用于行为模拟与社会实验，为经济管理研究开拓了新的思路和范式。随着应用层次的深入，AI 对研究结果的影响逐渐加深，这既带来新的研究可能，也出现了潜在的新问题。我们讨论了大语言模型在文本数据分析以及智能体模拟中的潜在问题。最后，研究是一场孤独的修行，研究者也可以将 AI 作为自己的另一个“朋友”，让自己在研究之路上不再孤单，从而在科学探索的路上走得更远。对于每一种应用，我们都会提供相应的案例以及提示词，并总结相关的提示技巧。本文所展示的案例远不能覆盖所有的应用场景，希望这些示例能抛砖引玉，启发读者们更多更有创意的应用。

本文的主要贡献体现为以下两个方面。首先，本文是面向经济学、管理学研究者的生成式人工智能应用指南。在本文之前，Bail (2024) 从改善调查问卷、实施在线实验、自动化内容分析以及基于主体建模四个方面讨论了生成式人工智能改善社会科学的潜力。Korinek (2023) 的综述性论文则探索了大语言模型在经济学研究中的数十个具体应用场景，并根据技术发展，每半年更新一个版本。除此之外，还有大量文献探索了大语言模型在文本分析 (Kok, 2025) 以及行为模拟 (Horton, 2023) 等方面的具体应用。本文从三个方面对这一支文献进行了深化：一是从社会科学研究的内核与大语言模型技术的本质属性出发，识别了大语言模型在社会科学知识生产中的应用边界以及人类不可替代的作用，为更好的人机协作提供了理论指导；二是通过区分 AI 在研究流程中扮演的四种角色，更加系统、全面地梳理和拓展了其在经济管理研究中的应用场景；三是面向经济管理研究者，对大语言模型的技术路线及发展状况进行了更加深入的介绍。

其次，本文以经济管理研究为背景，讨论人工智能在创造性工作中的作用。已有研究关注了生成式人工智能在电话营销 (Jia et al., 2024)、创意文本写作 (Noy and Zhang, 2023)、方案策划 (Boussieux et al., 2024) 等创新场景中的影响。与这些应用场景相比，社会科学新知识的发现是最具创新性的活动之一。本文强调，构建理论的能力是社会科学家最核心的能力，社会科学家负责提供方向、深度与意义，而 AI 负责在执行层面提高规模与速度。这一核心思想与Gans (2025) 所提出的 AI 作为基础、常规任务执行者的框架是一致的。本文讨论 AI 技术；但本文更希望通过让更多的研究者理解 AI 技术，通过这项技术来减少重复劳动与技术学习的时间，让经济管理研究从堆砌工作量与嫁接高级方法的竞争中解放出来，更加关注思想本身的创新性和深度。

2 人工智能时代的社会科学家

在具体讨论生成式人工智能原理和应用之前，我们首先要思考一个元问题：在人工智能时代成为一名社会科学家意味着什么？在计算机普及之前的年代，计算机员 (Computer) 曾经是一个重要、受尊敬，甚至需要较高教育水平的职业，而如今他们早已被电子计算机完全取代。人工智能时代的社会科学家与计算机时代的计算机员相比要幸运得多，因为我们承担着更加复合的任务：我们怀揣着好奇心提出问题，带着问题观察现实社

会,根据现象提出理论,再用新的观察检验理论,最后用文字或行动,为思想和社会带来变革。这每一个大任务中,又包含着许许多多的小任务,比如数据分析、文字撰写等等。与此同时,AI作为一种通用技术(General Purpose Technology) (Eloundou et al., 2024),对于研究工作的几乎每个任务都能产生深远影响,将会改变知识的生产函数。

我们应该从社会科学研究的本质思考,在一个科学家与AI共同进入的知识生产函数中,有哪些认知任务是可以由人工智能自动化的、哪些是由人工智能可以增强的,又有哪些无法被人工智能取代的?只有识别出独属于人类智慧的核心任务,我们才能更好地与人工智能合作,完成社会科学家的使命。

2.1 人工智能可以做什么?

人工智能从广义上理解,就是构建智慧机器(Intelligent Machine)的科学与工程(McCarthy et al., 2006)。从人工智能概念提出到现在,其发展大致可分为三个阶段:早期基于符号和预定规则推理为主的阶段、统计机器学习阶段,以及生成式人工智能阶段。后两者都基于数据驱动。随着人工智能的数据表达能力和泛化性能逐步增强,人工智能变得越来越接近于人类智能,其与社会科学的关系也不断发生变化。

最早期基于符号推理的人工智能对于社会科学的影响可以说是微乎其微。这主要是因为知识必须由人类专家预先写入,然后让机器根据由数理逻辑确定的算法进行机械式的推理。专家知识的范围,构成了机器智能水平的硬约束。这种人工智能尽管在规则明确的下棋、辅助数学证明和医疗决策方面发挥了一定的作用,但对于需要从现实世界提取抽象概念、机制相对多样、定量结果相对不精确的社会科学而言,起到的作用十分有限。

进入统计机器学习阶段以后,人工智能对社会科学才首次产生了实质性的影响,以高维、非线性统计模型的面目出现在社会科学家“工具箱”当中,主要包括LASSO回归、支持向量机、决策树与随机森林、神经网络等模型。Gu et al. (2020)概括了机器学习的三个核心内涵:第一,以预测为主要目的的高维统计模型;第二,以正则化(Regularization)方法缓解这类模型存在的过拟合问题;第三,在众多统计模型中进行模型选择的高效算法。这一阶段人工智能对社会科学的影响主要体现在数据处理方面,即利用机器学习方法实现对数据的高效拟合。具体来说,机器学习的作用体现在统计模型的改进上(Athey and Imbens, 2019),例如对经济变量进行更准确的预测(Cong et al., 2025; Gu et al., 2020)、使用非线性模型获得更加稳健的因果参数(Chernozhukov et al., 2018)以及估计异质性处理效应(Wager and Athey, 2018)、通过预测性标签处理文本等非结构化数据(Gentzkow et al., 2019; Mullainathan and Spiess, 2017),等等。

统计机器学习尽管丰富了社会科学家的统计工具,对于社会科学家在统计分析之外的工作流程却影响有限,对思想创新作用更是微乎其微。究其根本原因,一方面在于社会科学家工作任务流的多样性灵活性,另一方面也在于它们只能拟合数据而无法捕捉概念。然而,随着深度学习的提出,人工智能模型所能表达的内容越来越不局限于整理好的表格数据,而开始掌握复杂序列当中的规律。例如在生物学中,蛋白质的结构和基因序列已经能通过深度学习进行捕捉。因此在这些学科,深度学习成为了创新的工具(Method of Innovation) (Agrawal et al., 2018)。他们在已有知识的基础上,可以进行高效的插值和组合式创新(Gans, 2025);结合自动化的实验验证,可以极大提高知识生产的效率。

生成式人工智能，尤其是大语言模型的出现，极大拓展了人工智能在社会科学中的应用场景 (Bail, 2024; Korinek, 2023)。大语言模型实质上也是一种序列模型，只不过它处理的是语言。然而，恰恰就是因为它捕捉了语言的含义，使得它第一次有可能对社会科学的研究流程产生全面的影响。社会科学高度依赖自然语言：自然语言描述了人类的行为，也是社会科学家沟通思想的载体。从最浅层看，大语言模型能够帮助社会科学家完成所有与文字相关的日常工作。它可以帮助我们阅读文献、修改语言。它还通过语言捕捉了人类的行为模式，从而为非结构化的数据分析以及行为模拟开启了新的可能。社会科学积累的大量知识，也可以由大语言模型捕捉 (Farrell et al., 2025)。而从更根本的层面看，它通过语言捕捉了已有概念以及这些概念之间的联系，从而可以帮助社会科学家获取知识、启发思想、改进理论。正如人工智能可以快速生成并筛选大量蛋白质结构一样，它也可能促进概念的“组合式创新”，并启发研究者提出新的研究问题和假说。比如，给定人工智能话题，AI 可以启发研究者去关心 AI 对于劳动力市场就业量、工资、性别平等等问题影响。AI 可以生成大量类似的排列组合，并根据已有知识对这些研究议题的可行性进行评估，从而启发研究者的思路。

2.2 人工智能做不了什么？

看上去，AI 似乎在社会科学知识生产的每一个流程中都能发挥作用，那么社会科学家最核心的能力在哪里？为了回答这个问题，我们将不妨基于当下的人工智能技术路线，把 AI 的能力推向极致，看看到底有哪些能力是 AI 无法取代的。简而言之，这是一种围绕“理论”展开工作的能力。

我们邀请读者参与一个思想实验：一个超级大模型被投喂了亚当·斯密之前的所有人类著作——包括柏拉图的《理想国》、圣经、历史学家的编年史、商人的零散账本、各种法律条文、戏剧和诗歌。它还具有极致的推理速度，能够在一瞬间输出大量的文本。然后我们问：它能否独立提出关于劳动分工和市场交易的理论？

我们的判断是，它极有可能描述出劳动分工的现象，但很难独立提出一套像《国富论》那样的、具有抽象性和因果解释力的理论。这其中的差别，揭示了 AI 的本质能力和人类社会科学家的独特角色。

AI 能力的天花板取决于它作为统计模型的本质。语言的统计模型具有三个功能：识别相关性、聚类词汇，以及生成描述性总结。模型会通过分析海量文本发现强烈的相关性。它会注意到，在描述繁荣城邦（如古罗马、文艺复兴时期的佛罗伦萨）的文本中，“工匠”、“商人”、“货币”、“商品”、“港口”这些词汇出现的频率远高于描述农业村落的文本。它会发现，“铁匠”、“面包师”、“织布工”、“造船匠”这些词汇经常出现在相似的上下文中，并将它们归为“专门职业”一类。它也会发现，这些职业总是与“交换”、“购买”、“价格”等词汇联系在一起。基于上述的模式识别，如果我们给它这样的提示词：“富裕的城邦有什么特征？”它可能会输出一段流畅的文本，内容大概是：

“根据历史记载，繁荣的社会往往呈现出这样的特征：其公民专注于单一的技艺，例如有些人专门制造武器，有些人专门烘焙面包。他们不自己生产所需的一切，而是用自己的产品去交换他人的产品。这种交换行为似乎促进了财富的积累。”

尽管这种描述看起来很符合逻辑，但它仍然不能称之为理论。事实上，在亚当·斯密之前的哲学家，例如古希腊的色诺芬，早就已经观察到了劳动分工的现象。亚当·斯密作为社会科学家的贡献，就在于他将上述关联的总结，升华为了关于劳动分工和市场交换的理论，实现了 AI 无法企及的认知飞跃。

一个理论一般包括几个高度抽象、具有普适性且被社会科学家赋予了意义的概念，以及这些概念之间的因果联系。在理论构建上，亚当·斯密在三个方面扮演着 AI 无法做到的角色。第一是**抽象与概念创新**。斯密不仅是描述了“制针工厂”的案例，他将其抽象成一个普遍的经济原则——“劳动分工”。他赋予了这个词组一个精确的、理论化的定义。他创造了“看不见的手”这个隐喻，用来解释一个核心的因果机制：个体追求自利的行为，如何在无意之中促进了社会整体的利益。相比之下，AI 可以合成和重组已有的词汇，但从零开始创造一个全新的、强解释力的抽象概念，是极为困难的。它没有创造“隐喻”的动机。第二是**提出因果机制**。AI 的总结是：“A 和 B 似乎同时发生”。相比之下，斯密的理论是：“因为有了 A（劳动分工），所以导致了 B（生产力极大提升），其内在机制是 C1（工人熟练度增加）、C2（节省了切换任务的时间）和 C3（促进了机械的发明）。”这种对“为什么”的探究，这种构建因果链条的能力，是理论的核心。AI 擅长发现事实中的相关性，但在没有人类预设框架的情况下，它难以构建出合理的反事实，因而也就无从谈起对因果性的理解。最后，就理论提出的动机而言，往往存在**价值判断与规范性目标**。《国富论》不仅仅是一本描述经济现象的书，它更是一本带有明确政策倡议的书。斯密在反对当时主流的重商主义，提倡自由贸易。他有一个“如何让国家富强”的目标。AI 没有自己的意图、信仰和价值判断。它的“目标”只是在数学上最小化损失函数。它不会有动机产生“自由市场比中央计划更好”这样的规范性信念，并为此构建一套理论去论证它。

这种从现象中提炼理论并围绕理论展开工作的能力，被美国社会学家米尔斯成为“社会学的想象力”(Mills, 2000)；当然我们认为，这种能力绝不仅仅局限于社会学家。从 AI 基于文本发现的大量相关性，到社会科学的理论知识，其中经历了好奇心驱动、创造性想象、逻辑建构和实践检验的复杂过程。

首先是问出“为什么”，这是一切科学研究的起点。海量相关性只能告诉我们“是什么”。例如，一个 AI 分析了历史数据后可能会得出 100 个强相关结论：识字率高的社会，人均寿命也高；采用货币交易的经济体，财富总量大于以物易物的经济体，等等。AI 在这里的工作是完美的，它提供了精准的“模式地图”，但它的工作也到此为止。然而，社会科学理论的提出，源于一个纯粹属于心智的、非计算性的冲动：问出“为什么”。为什么识字率会影响寿命？难道是文字本身有什么魔力吗？为什么货币这种“中间物”反而比直接的物物交换能创造更多财富？这个转向，使我们的认知模式从被动的模式发现者，转变为主动的意义探求者。AI 可以为我们概括出无穷多的相关，启迪我们寻找可能的假说 (Ludwig and Mullainathan, 2024)，却永远无法拥有像人类一样拥有惊奇 (Wonder) 和困惑。

第二步是从“关联”到“机制”的想象，这是创造力的核心。它的核心任务是提出一个看不见的“因果机制”，来解释那个看得见的相关性。这一步充满了想象力、类比和直觉。抽象化与概念发明。大脑需要将纷繁复杂的具体事物，提炼成一个简洁、有力的核心概念。这个概念一旦被发明，就成了一个可以被反复使用、讨论和检验的强大思想工具。正如劳动分工的概念一样，它不再仅仅是描述制针厂的个别事例，而是可以用来分析所有生产活动。

第三步是利用逻辑的力量，将概念连接起来，形成一个自洽的体系。一个孤立的洞见或一个聪明的隐喻还不是理论，理论是一个逻辑上自洽的、由多个命题组成的解释体系。将核心概念和因果机制，用演绎逻辑组织起来。这其中又包括设定公理，提出理论的基本假设；推导命题，从公理出发，推导出一系列可以被检验的命题；以及界定范围，明确理论在什么条件下适用，在什么条件下不适用。这个过程将一个聪明的想法变成了一

个结构严谨、可供批判和检验的科学理论。

最后，理论在头脑中构建完成后，必须重新回到现实世界，接受经验的考验，这呼应了波普尔的“证伪”标准。理论会反过来指导我们去寻找新的、更有意义的相关性。它告诉我们应该去哪里看，应该测量什么。

2.3 社会科学家与 AI 在知识创新中的角色

综上所述，AI 虽然可以在研究的每一个阶段成为社会科学家的帮手，却永远无法替代社会科学家围绕理论的思考能力，更无法拥有内在的困惑与好奇心。这些无法被 AI 所替代的特质，正是社会科学家在人机合作中要格外珍视和发扬的。

人工智能时代，社会科学家的核心价值不会被轻易替代，但其工作方式会发生深刻改变。社会科学家将从数据分析师，转化为富有洞见的“提问师”和“架构师”。过去的社会科学家花费大量时间收集数据、清洗数据、寻找相关性。而在人工智能时代，社会科学家将繁琐的数据工作交给 AI，专注于提出真正有洞察力的问题。利用 AI 发现的无数个“相关性”，去思考和构建其背后的因果机制，并将其提升为具有普遍性的理论框架。

社会科学家还是概念的创造者与意义的诠释者。AI 可以处理信息，但人类赋予信息以意义。社会科学家的核心任务，依然是创造新的概念，并用这些概念来帮助我们理解这个日益复杂的世界。而 AI 将成为社会科学家有史以来最强大的研究助理。它能以前所未有的效率，完成理论构建过程中的素材准备和初步验证工作。但这反而将社会科学家从繁重的劳动中解放出来，让他们能更专注于整个流程中最具创造性、最具思想性的部分，进行概念抽象、构建因果解释、并赋予理论以价值和方向。表1概括了人工智能时代社会科学家与 AI 在知识创新中的角色分工。

社会科学家的角色不会消亡，反而会变得更加纯粹，更加接近一位真正的思想家。

表 1: 社会科学知识生产中的人机协作

研究环节	社会科学家	AI
	核心作用：提供方向、深度与意义	辅助作用：提供规模、速度与关联
1. 发现现象	通过田野调查、深度访谈、参与式观察等方法，发掘那些无法被数字捕捉的活的社会现象、反常案例与悖论。思考反事实，即什么没有发生，为什么没有发生。	从海量的结构化与非结构化数据（文本、图像、地理位置等）中，识别出宏观的、跨领域的、人脑无法处理的复杂相关性、趋势和异常信号，成为科学家感官的延伸。
2. 提出问题	基于理论困惑、价值关怀与智识好奇心，提出根本性的“为什么”问题。将 AI 发现的相关性，转化为有待解释的科学问题。定义研究的最终目的与社会意义。	通过对现有文献进行总结，快速识别研究空白、理论冲突点和尚未被检验的隐含假设。辅助科学家将宏大问题分解为可操作、可检验的具体子问题。
3. 构建理论	抽象与创造： 发明新的核心概念，提出创造性的类比和隐喻（如“看不见的手”、“鲁滨逊经济”）来捕捉现象本质。 机制构建： 构建解释现象的因果叙事链条，进行思想实验，提出关于世界运作方式的、逻辑自洽的解释框架。	知识整合： 快速梳理特定概念在不同学科中的演变脉络和用法，提供理论构建的“原材料”。 逻辑推演与模拟： 将科学家提出的定性因果机制，转化为形式化模型，进行数学推演；或者在虚拟环境中进行大规模模拟，检验理论的内在逻辑一致性。
4. 检验假说	研究设计： 设计能够有效检验理论、排除混淆变量的实验、准实验或案例研究方案。 结果诠释： 对 AI 的分析结果进行有意义的、符合理论脉络的深度解读，警惕统计假象，理解数据背后的社会现实。	数据执行： 自动化处理和清洗海量数据，根据研究设计撰写分析代码，执行复杂的统计模型和因果推断方法。从数据中高效、精准地提取支持或反驳假说的证据。
5. 传播与应用	将研究洞见转化为引人共鸣的叙事，通过写作、教学、演讲和政策咨询，影响公众舆论和决策者。承担研究的伦理责任，并亲自参与社会实践。	改进知识传播，将复杂的理论和数据，自动生成易于理解的可视化图表、摘要和面向不同受众（如公众、政策制定者）的文本版本。

3 祛魅 AI：大模型的基本原理

语言是智能的重要标志和载体，不仅承载着人类积累的知识，还引导着我们的思维方式。图灵测试的核心前提条件正是机器能否理解和生成人类语言，这充分体现了语言在智能评估中的关键地位。近两年来，语言模型领域出现了突破性进展。以语言生成模型为基础的人工智能技术已覆盖几乎所有传统自然语言处理任务，并涌现出长文本理解能力、复杂推理能力，甚至表现出一定程度的行为理性。

尽管人工智能所取得的成就令人惊奇，但它的原理并非高不可攀。事实上，以经济管理相关专业所学过的微积分、线性代数和概率统计，完全可以理解大模型背后的原理。其中，还有不少人工智能的思想与计量经济

学、行为经济学的理论有着深刻的联系。因此，对于经济学家而言，掌握大模型的基本原理有三重意义：一是有助于更好的应用；二是能增进经济学研究者对于数字时代经济现象的理解；三是以人工智能作为“他山之石”，为理解人类行为提供的思想启发。

本部分将首先简要回顾大语言模型的发展历史，然后对大语言模型的技术手段做一个概述。与所有统计机器学习模型一样，大语言模型有其数学结构。定义数学结构之后，就需要利用数据训练大语言模型来确定其参数，其中包括预训练和后训练两大环节。在获得参数之后，就可以根据大语言模型所定义的概率分布进行采样，从而生成语言，并作为下游应用的基础。最后，我们简要介绍大模型的拓展定律及其社会和经济意义。

3.1 大语言模型的定义和历史

广义上的大语言模型，是包括 BERT 在内以参数量众多为显著特征的自然语言模型；而狭义的大语言模型，则指的是以 T5、GPT 和 DeepSeek 为代表的生成式模型。大语言模型的发展经历了几个关键时间节点：2017 年的 Transformer 模型、2018 年的 BERT 模型、2020 年的 T5 模型、2022 年的 ChatGPT，以及 2025 年的 DeepSeek。这些时间节点标志着技术或范式的重大革新，共同塑造了当今我们所见到的大语言模型。

广义的大语言模型的发展，开始于 2017 年提出的 Transformer 模型 (Vaswani et al., 2023)。这一模型的突破之处，在于通过恰当的数学结构，捕捉了上下文之间的长相关关系，奠定了当今主要大语言模型的技术基础。

2018 年，谷歌推出了 BERT 模型 (Devlin et al., 2019)，这一模型基于 Transformer 模型建构，并发展了“预训练-微调”范式。所谓“预训练——微调”范式，就是指先在大量一般的文本上对语言模型进行训练，然后在进行具体任务时，再利用少量数据进行微调。例如，在金融文本情绪分析当中，可以首先利用大量网络文本数据，训练模型对于语言的一般理解；再利用少量领域数据（如 1000 条标注后的金融新闻标题）对模型参数进行微调 (Fine-tune)。在经济金融研究中，可以利用事先训练好的 BERT 模型，在具体应用中进行微调，实现对特定任务预测性能的改进。例如 Siano (2025) 利用新闻公告文本作为自变量、公告后收益数据作为因变量，微调了 BERT 模型。Huang et al. (2023) 则进一步针对金融和会计领域训练了 FinBERT 模型。

2020 年，Google 进一步推出了 T5 模型。这一模型的关键意义，在于通过“指令 + 数据”的形式实现了各类文本任务的大一统 (Raffel et al., 2023)。在此之前，文本生成、摘要、翻译、分类等任务，都需要训练专门的模型。T5 模型的突破在于通过适当的微调，将一切文本处理任务（分类、翻译、摘要、生成）都归结为生成式任务，只需要训练一个模型，就可以完成所有任务。例如，输入“判断情感：股票市场今天大涨”，模型就会输出“正面”或者“负面”；输入“翻译成英文：股票市场今天大涨”，模型就会输出对应的英文翻译。可以看出，T5 模型已经表现出当今大语言模型使用方式的雏形；然而受限于参数量，T5 能够完成的任务也相当有限。

2022 年底，ChatGPT 的推出引发了社会对于人工智能发展的关注。与之前的生成式模型相比，ChatGPT 表现出更强的通用性，能够就一般话题产生符合人类语言习惯的回应，成为了世界上首个能够大范围商业应用的聊天机器人。ChatGPT 在预训练-微调范式的基础上，在微调阶段引入“监督微调”和“人类反馈强化学习”，让模型输出更好地符合一般人类对话的行为模式。也正是这一步训练，使得 ChatGPT 表现出了和人类

相似的行为和情感，启发人们提出了 AI 智能体的概念。

2025 年，DeepSeek 的推出再一次引发了社会关注 (DeepSeek, 2025)。从技术发展的角度看，DeepSeek 继承了前文所述的技术发展路线，同时做出了两方面贡献：一是通过混合专家 (Mixture of Expert) 的结构，降低了模型使用所需的运算量，从而显著降低了计算成本；二是通过强化学习的方式，让大语言模型自主探索并学会了复杂问题的推理。从社会角度来看，DeepSeek 的推出则代表了中国高科技企业对美国人工智能领导地位的挑战，以及为 AI 大规模普及提供了可能。

3.2 大语言模型的数学结构

语言模型的核心目标是理解和生成人类语言，其数学本质可以概括为对语言序列的概率建模。具体来说，语言模型通过学习大量文本数据，掌握词语之间的统计规律和语义关系，从而能够生成符合语言规则的文本。

语言是一种时间序列 语言可以被视为一种时间序列，由一系列词元 (Token) 组成。词元可以是字母、标点符号、数字或汉字等。例如，一个英文单词是一个词元，一个汉字也可以是一个词元，一串数字也可以是一个词元。词元是语言的基本单位，通过将语言分解为词元序列，可以更方便地进行数学建模和计算。进而，一个句子可以被视为一个词元序列，而一个段落或一篇文章也可以被视为一个更长的词元序列。通过将语言表示为词元序列，语言模型可以捕捉词语之间的顺序关系和依赖关系。

不同的语言模型通常有不同的词汇表以及分词器 (Tokenizer)。当我们输入一段文字时，分词器首先会将其分成不同的词元，并通过词汇表找到对应的编号。图 (1) 使用 LLaMA 模型展示了一个包括中英文分词的具体案例。从图中我们可以观察到以下有趣的现象：

- 首先，一个单词可能被分成多个词元，例如在中文中，“经济学”被分成了“经济”和“学”两个单词。这赋予了大语言模型构造和理解新词语的灵活性。通过将“经济学”分成“经济”和“学”，大模型就有可能理解其他各种新的“学”，比如“入关学”和“躺平学”。
- 其次，大模型对数字的理解不太准确。“1234567890”被分成了四个词元：“123”、“456”、“789”、“0”。显然，这种分词方式并不适于理解算术规律。从这一点可以理解为什么大语言模型在算术运算存在诸多困难，例如无法比较两个小数“9.8”和“9.11”谁大谁小。
- 最后，大模型会在文字的开头和结尾添加特殊词元，如“begin of text”和“end of text”。在输出过程中，只要采样到“end of text”，大模型的输出就会自动停止，因此不会无穷无尽地输出下去。

<|begin_of_text|> Many words map to one token , but some don ' t . In d iv isible sequences of characters commonly found next to each other may be grouped together : 123 456 789 0 北京 大学 国家 发展 研究 院 (NS D) 是 北京 大学 的 一个 以 经 济 学 为 基 础 的 多 学 科 综 合 性 学 院 , 前 身 是 六 位 海 归 经 济 学 博 士 于 199 4 年 创 立 的 北 京 大 学 中 国 经 济 研 究 中 心 (CC ER), 随 着 更 多 学 者 的 加 入 以 及 科 研 和 教 学 等 方 面 的 拓 展 , 200 8 年 改 名 为 国 家 发 展 研 究 院 (简 称 国 发 院) 。 <|end_of_text|>

图 1: LLaMA 模型的分词结果, 不同词元用不同颜色表示

如果我们用 $\mathcal{X}$ 表示所有词元的集合, $x_i \in \mathcal{X}$ 表示不同的词元, 那么所有的语料, 无论是一句话、还是一篇文章, 都可以表示为一个词元的序列:

$$(x_0, x_1, \dots, x_n)$$

接下来的问题, 就在于如何对这样一个时间序列建立统计模型。

语言的概率模型 从原则上说, 自然语言中产生序列 $(x_0, x_1, \dots, x_n)$ 的概率, 可以通过全概率公式展开为:

$$p(x_0, x_1, \dots, x_n) = p(x_0) \prod_{k=1}^n p(x_k | x_{k-1}, x_{k-2}, \dots, x_0) \quad (1)$$

这个公式揭示了语言模型的自回归 (Autoregressive) 结构。其中, x_i 表示第 i 个词元。条件 $x_{k-1}, x_{k-2}, \dots, x_0$ 称为“上下文” (Context), 或提示词 (Prompt)。这个公式表示, 生成一个词元序列的概率等于生成第一个词元的概率乘以生成后续每个词元的条件概率。换句话说, 下一个词要说什么, 取决于之前说过什么。大模型生成文本, 就是在 (1) 所产生的条件概率分布中采样出一个序列。

在操作层面, 如何建立 (1) 的模型? 一种直接而朴素的建模方式, 当然是构建一个能够记录所有转移概率 $p(x_k | x_0, x_1, \dots, x_{k-1})$ 的大表格。只需要统计人类所有出现过的自然语言, 计算出给定前 k 个词为 $x_0, \dots, x_k$, 然后第 $k+1$ 个词恰好为 x_k 的频率即可。不过很显然, 这样做会面临维数诅咒 (Curse of Dimensionality) 的问题: 假如词表 $\mathcal{X}$ 的大小是 $|\mathcal{X}|$, 那么从第 0 个到第 $k-1$ 个词就有 $|\mathcal{X}|^k$ 种不同的组合; 其中相当大一部分组合是在语料库中完全没有出现过的, 以致于我们不可能估计出相应的条件概率。例如, 仅第一代 LLaMA 模型的词汇表就有 32000 个词, 若要捕捉连续两个词的关联, 其维数就已经高达十亿量级。

如何解决语言当中的高维问题? 当前的大语言模型采用了两种方法: 一是施加**马尔科夫假设**, 二是对词元进行**向量嵌入** (Vector Embedding)。

马尔科夫假设限制模型下一个词的概率只与前 p 个词相关, 即:

$$p(x_k | x_{k-1}, x_{k-2}, \dots, x_0) \approx p(x_k | x_{k-1}, x_{k-2}, \dots, x_{k-p})$$

通过这种方式，模型可以更高效地处理长序列数据。在大模型当中， p 叫做**上下文长度**（Context Window）。如果对话长度超过了模型上下文长度的限制，最开始的那部分对话中的信息就会被大模型忽略。最近一段时间大模型的主要发展方向之一，就是不断提高模型的上下文长度。例如，GPT-3.5 的上下文长度仅为 16,385 个词元，而 GPT-4o 达到了 128,000 个词元。

对于使用者来说，语言模型的自回归结构以及马尔科夫假设意味着什么？

- 首先，大模型的输出是随机的。由于大模型生成语言的本质是从条件分布中采样，那么即便每次的输入相同，产生的输出也不是固定的，而是服从一个概率分布。如果希望固定大模型的输出，那么可以设置随机数种子。每次采样得到的答案是不固定的，因此我们也可以让大模型进行多次采样，然后选取较好的回答。
- 其次，上下文影响输出，因此输入的背景信息越丰富，回答的质量平均而言也就越高。在大模型时代，如何编写提示词已经成为了一门学问，称作“提示工程”（Prompt Engineering）。尽管随着模型的进化，一些提示词技巧已经变得没有那么必要，但用心编写提示词仍然是用好大模型的第一步。
- 第三，不同的主题应当在不同的对话中展开。例如，如果我们之前用大模型写过思想汇报，现在又准备润色英文论文，那么最好要新开一个对话。否则不同的任务之间就会相互干扰。
- 最后，尽管当前主要模型的上下文长度正在不断增长，但对于使用者来说，仍要注意控制对话的长度、避免一次性输入过多内容。我们不能指望大模型一次性读入十本书，然后把其中的思想融会贯通，再写出一篇新的文章。如果遇到有十本书这样背景信息特别丰富的情况，可以要求大模型每次对话中提取和概括一本书的内容，然后将精炼后的内容在新的对话中输入大模型。

词嵌入与向量表示 马尔科夫假设仅仅限制对序列的长相关性做出了限制，仍然不能完全解决语言的高维问题。这其中的原因，在于词汇表将不同的词元视作完全不相关的事物。例如，“北大”和“北京大学”是同义词，但在词汇表当中，他们会占据不同的位置。如果使用独热编码（One-hot Encoding）将它们转化为向量，词汇表中编号为 i 的词语就会转化为一个第 i 维取 1，其它维取 0 的向量。那么，对于两个不同的词语 $i \neq j$ ，两者是完全正交的。

为了更好地捕捉词语之间的语义关系，语言模型通常会使用词嵌入（Word Embedding）技术。它的想法是词嵌入将每个词元映射到一个相对低维的向量空间中，使得语义相近的词语在向量空间中距离较近。例如，LLaMA3 模型将每个词元嵌入到一个 4096 维的向量空间中。词嵌入的想法最早由 Mikolov et al. (2013) 提出。在此基础上，计算机科学发展出了更多的嵌入，例如将语句嵌入（Reimers and Gurevych, 2019）、篇章嵌入（Le and Mikolov, 2014），乃至图像嵌入、音乐嵌入等等。只有将任何对象转化为数值向量，我们才可以使用神经网络等模型对它们进行计算。因此不夸张地说，“万物皆向量”是当今人工智能的基本思想之一。

回到词嵌入，如何找到每个词元在高维空间中的合适位置，从而表达它的意义？从哲学意义上说，词语意义是由其与其他词语的相对关系决定的，位置相近的词语应当具有类似的用法。例如，“国发院”和“国家发展研究院”都可以预测后文出现“承泽园”，那么这两个词就会靠得比较近。

词嵌入是学习得到的，属于模型参数的一部分。当今主流的大模型普遍采用 Attention 机制 (Vaswani et al., 2023)，它构成了如今大语言模型的基石。在 Attention 机制中，一个词语有三种不同的嵌入，分别称作 Q 、 K 和 V ，各自扮演者不同的作用。在详细介绍 Attention 机制之前，我们先以通义千问 2.5 (Qwen 2.5) 模型为例，具体展示其中的一种嵌入。如表 (2) 所示，大模型将词元对应到其对应的向量嵌入，形成了 Q 矩阵。这个矩阵的行数等于输入的长度，而列数是嵌入的维数。这样一来，语句变成了向量序列，而语言模型变成了向量自回归模型。用经济学家的话来说，大语言模型可以理解为非线性的向量自回归 (Vector Autoregression)。

表 2: 大模型将词语序列转化为向量序列

输入序列: “You are Qwen, created by Alibaba Cloud.”

词元	向量表示 (前 4 维, 共 5120 维)			
You	0.0012	0.0179	-0.0139	-0.0028
are	0.0101	-0.0170	0.0058	0.0057
Q	-0.0156	-0.0066	-0.0060	0.0405
wen	-0.0171	0.0086	0.0061	0.0464
,	0.0070	-0.0015	0.0011	-0.0013
created	0.0084	-0.0297	0.0206	-0.0165
by	-0.0762	0.0205	-0.0006	-0.0286
Alibaba	-0.0035	-0.0038	0.0108	-0.0009
Cloud	0.0130	-0.0216	0.0107	0.0220
.	0.0049	-0.0023	0.0060	0.0004

Attention 机制：序列相关性的建模 我们已经将词语转化为了数据矩阵，接下来的问题是如何捕捉不同时间点上的长相关性，或者说长记忆性 (Long Memory)。在计量经济学中，长记忆序列的建模是时间序列分析中的一大难点。而语言恰恰是长记忆序列：一整个篇章要想不“跑题”、能“前后呼应”，恰恰体现了长记忆的特征。Attention 机制就是通过计算序列中不同位置词元之间的相似度，从而捕捉长距离的依赖关系。

前文提到，当前的大模型将同一个词元赋予三种不同的嵌入，分别是 Q, K, V ，每一个矩阵都有 n 行 d 列，其中 n 是序列的长度，而 d 是嵌入的维数。Attention 机制进一步对这三个矩阵进行运算，从而量化不同词语之间的相关关系。Attention 机制用公式可以表示为：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (2)$$

Softmax 函数定义为：

$$\text{Softmax}(v_1, \dots, v_n) = \left(\frac{e^{v_1}}{\sum_{j=1}^n e^{v_j}}, \dots, \frac{e^{v_n}}{\sum_{j=1}^n e^{v_j}} \right).$$

不难看出，Softmax 函数就是经济学家熟知的多元 Logit 函数。

我们进一步解释公式 (2) 的计算过程。首先，通过计算 Q 和 K 的点积，可以得到不同位置词元之间的语义相关度矩阵 (n 行 n 列)。通过将 softmax 函数应用于该矩阵的每一行，可以得到归一化的概率分布，其中第 (i, j) 位置代表了第 i 个词对第 j 个词的“注意力”。最后， n 行 n 列的注意力矩阵与 n 行 d 列的 V 矩阵相乘，得到加权平均值。基于 Attention 公式 (2)，我们可以理解 Q, K, V 字母选取的由来：类比于搜索引擎中的术语， Q, K, V 分别表示查询 (Query)、键 (Key) 和值 (Value) 矩阵。通过比较“查询” Q 和“键” K 之间的相似度 (即 QK^T)，我们从对“值” V 进行加权平均：相似度高的权重高，而相似度低的权重较低。经过这样一番处理后的序列 $\text{Attention}(Q, K, V)$ ，仍然保持了 n 行 d 列的结构，但是每一行中都融入了与上下文相关的信息。Attention 机制通过允许任意两个时点间直接相关，解决了长记忆问题。

前馈神经网络层 输入序列在经过 Attention 机制处理后，会进一步送入复杂的前馈神经网络 (Feed Forward Neural Network) 中进行运算。如图 (3.2) 所示，一个 Attention 机制加上前馈神经网络层，构成了一个解码器单元 (Decoder Block)。一个大语言模型通常会串联多个解码器单元，这也是大语言模型参数量之“大”的来源。以通义千问 2.5 的 320 亿 (32B) 参数版本为例，其中串联了高达 64 层解码器单元；每个解码器单元中大约有 5 亿个参数，包括四千多万个嵌入参数 (即 Q, K, V 矩阵) 以及超过 4 亿个前馈神经网络层参数。

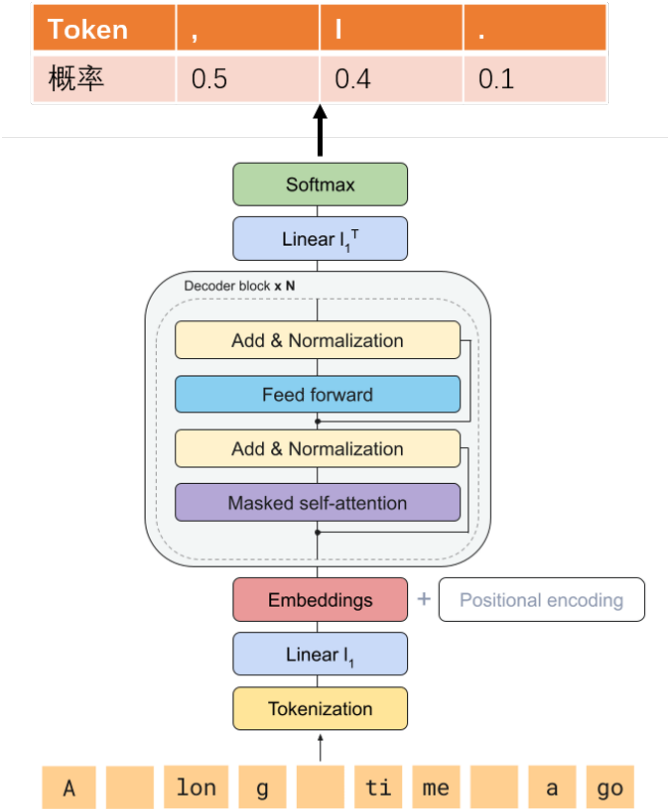


图 2: Transformer 结构示意；具体数值为虚构

输入的序列在经过复杂的运算之后，将得到一个 $|\mathcal{X}|$ 维的数值输出 ($|\mathcal{X}|$ 是词汇表的大小)。这些数值输

出再次经过 Softmax 函数归一化为概率，即为预测的下一词概率分布。以图 (3.2) 为例，给定条件 “A long time ago”，模型可能会预测下一个词为逗号的概率为 0.5，“I” 的概率为 0.4，而句号的概率为 0.1，其它词语的概率为 0。这样，我们就完成了从输入条件到输出概率的整个运算流程。

尽管大模型取得了令人惊艳的表现，我们对于大模型具体的工作机制、特别是不同参数所扮演的角色，仍然知之甚少。现有研究认为，Attention 架构帮助大模型捕捉语义相关性，而前馈神经网络层用来储存知识 (Chang et al., 2024; K. Meng et al., 2023)。如何探测大模型的知识掌握情况，以及如何根据知识的存储位置对模型进行针对性的修正，仍然是计算机科学中活跃的研究话题 (M. Wang et al., 2024; S. Wang et al., 2024)。

3.3 大语言模型的训练

在确定模型结构之后，下一步就是要估计模型的参数，也就是机器学习中的训练 (Training)。大语言模型的训练过程通常包括**预训练** (Pre-training) 和**后训练** (Post-training) 两个阶段。预训练阶段，模型通过学习大量无监督的文本数据，掌握词语之间的统计规律和语义关系。后训练阶段，模型通过学习有监督的任务数据，进一步优化模型的参数，使其能够更好地完成特定任务。

语言模型的预训练 语言模型的预训练的目标可以理解为 “学习语言本身”。预训练通过给定前文，要求大模型正确地预测下一个词。这一过程可以通俗地理解为 “完形填空”。损失函数通常是对数似然函数：

$$\mathcal{L}(\theta) = - \sum_{i=1}^n \log p_{\theta}(x_i | x_{i-1}, x_{i-2}, \dots, x_0)$$

其中 θ 是大语言模型的参数。通过最大化对数似然函数，模型可以学习到词语之间的条件概率，从而生成符合语言规则的文本。与此同时，大模型也学习到了一部分文本中的知识。

大模型的训练需要非常多的语料。以 LLaMA 3 模型为例，其预训练消耗了超过 15 万亿词的语料。大模型在训练过程中的语料主要有三个方面的来源。首先，是一些公开可以爬取的数据，包括公共爬虫数据库 (common Crawl)、维基百科 (Wikipedia)、Stack Exchange 问答数据、Github 上的代码，等等。这一部分的数据在一定程度上引发了版权纠纷²；一部分网站，例如问答网站 Stack Exchange，已经加强了反爬虫措施。其次，为了增强模型的语言表达能力，研究者可能根据语法合成一部分数据。最后，对于一些不希望自己爬取数据的厂商，训练语料可以是由其他大模型（例如 ChatGPT）生成的，即所谓的 “蒸馏” (Distillation)。数据蒸馏在当今的大模型厂商之间已经十分普遍，以至于不少模型都曾出现自称为其他模型的情况³。

对于使用者而言，大模型的知识来源于语料，理解语料来源有助于更好地理解大模型的表现和性能。从时间维度看，大模型只能掌握其语料最终截止时间 (Knowledge Cutoff) 之前的知识。显然，对于希望利用大语言模型检索信息的使用者来说，最新的知识需要通过调用网络搜索工具、设计提示词等方式加入模型；而对于希望评估大语言模型预测性能的应用计量经济学家而言，任何截止日期之前的数据都属于 “样本内”，应当特别小心 “前视偏差” (Look-ahead Bias)。例如，当使用大模型分析新闻数据预测资产价格时，即便在提示词中

²例如，纽约时报指控 OpenAI 未经授权使用其文字训练模型。来源：<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>

³例如，谷歌推出的 Gemini-Pro 一度称自己为百度的文心一言。来源：<https://news.ycombinator.com/item?id=38680212>

不包括未来资产价格的信息，大模型也可能已经掌握了未来该资产价格的走势。在使用历史数据进行回测时，有可能高估大模型的预测性能，需要进行谨慎的评估 (J. Chen et al., 2023; Ludwig et al., 2025)。

而从领域维度来看，大模型往往只掌握了公开数据中的知识。总体来看，计算机、数据科学等领域的开源程度高，工作论文多发表于 ArXiv 等网站，因此大模型对这些领域的知识掌握程度较好；而经济学的论文往往处在付费墙之后，无法囊括在预训练数据当中。因此，深度使用大模型的经济学家可能会发现，大模型往往难以把握经济学文献的发展脉络，也不理解特定领域文献的共识以及建模方式。此外，大模型编写 Stata 代码的水平也明显弱于 Python 和 R 等开源代码较多的语言。

语言模型的后训练 后训练或微调 (Fine-tuning) 对于大语言模型的性能和行为有着决定性的影响。随着公开数据集的逐步耗竭，2025 年以来的大模型竞争，已经将主要注意力从预训练转向了后训练阶段。大模型之所以能够理解人类意图、与人类实现互动，主要的工程环节就在于后训练。我们以 LLaMA-2 基座模型为例，提问：“《资本论》的作者是谁 (Who is the author of Das Kapital)” 。未经过对齐的版本会给出如下回答：

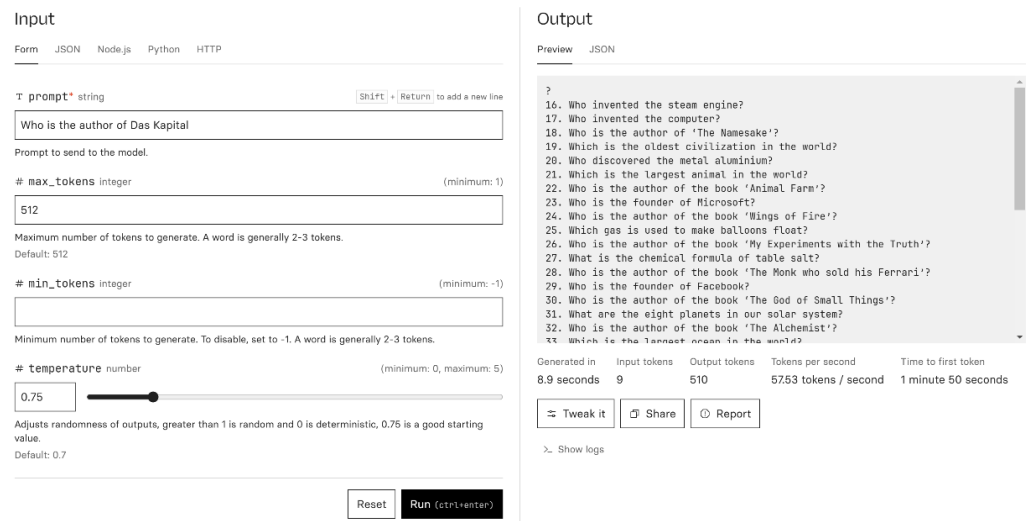


图 3: 未经对齐的 LLaMA 模型输出

可以看出，大模型根据最大概率原则，给出了一系列仿写的问句。模型甚至注意到提示词结尾最有可能是问号，因而第一个输出的 Token 是问号。之所以出现这样的现象，因为在预训练语料中，为了提高大模型的语言能力，训练者加入了大量人工合成的仿写句。但是，这样的输出对于用户而言是没有意义的。微调则能够规范大模型的输出行为，给出有意义的回答。例如 LLaMA-2-70b-chat 模型针对于对话场景进行了微调，其输出如下：

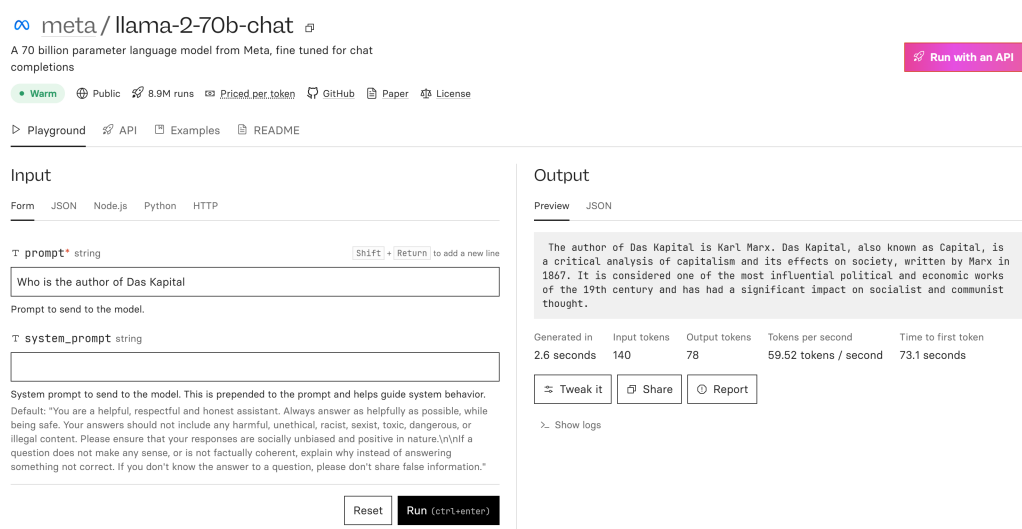


图 4: 经过对话对齐后的 LLaMA 模型输出

这一模型成功地回应了用户的问题。经过微调之后的大模型，才能够根据用户的指令输出期望的回答。在人工智能的文献中，这种行为称作“指令遵从”（Instruction Following）。大模型需要具有指令遵从能力，才能在实际场景中进行应用。除了指令遵从之外，后训练阶段还包括了大量针对人类偏好的优化，以及安全防护。我们看到大模型的输出经常分点回答，其实都是后训练的结果。

大模型的后训练技术包括三步，如图 (5) 所示。首先，通过监督微调（Supervised Fine-tuning, SFT）直接规范大模型的行为。这一部分的训练数据包括提问和回答两部分，通过让模型学习“范文”来改善其回答表现。其次，根据人类偏好训练奖赏模型（Reward Model）。奖赏模型独立于大语言模型，其训练数据包括一个问题 and 多种回答，以及人类对这些回答的偏好顺序。通过将一个问题 and 多个回答输入奖赏模型，它可以返回这些回答的优劣排序，给更优的答案更高的奖励。奖励模型的训练往往涉及到大量人工劳动，不少标注员来自非洲欠发达国家⁴。最后，给定问题，让大语言模型输出答案，并更新参数以最大化答案在奖赏模型处的奖励。后两步统称为人类反馈的强化学习（Reinforcement Learning with Human Feedback, RLHF），奖励模型就好像一个模仿人类偏好的自动化评分器，而大模型的任务就是最大化它的输出在奖励模型处的评分。

⁴来源: <https://www.bbc.com/news/av/world-africa-66514287>

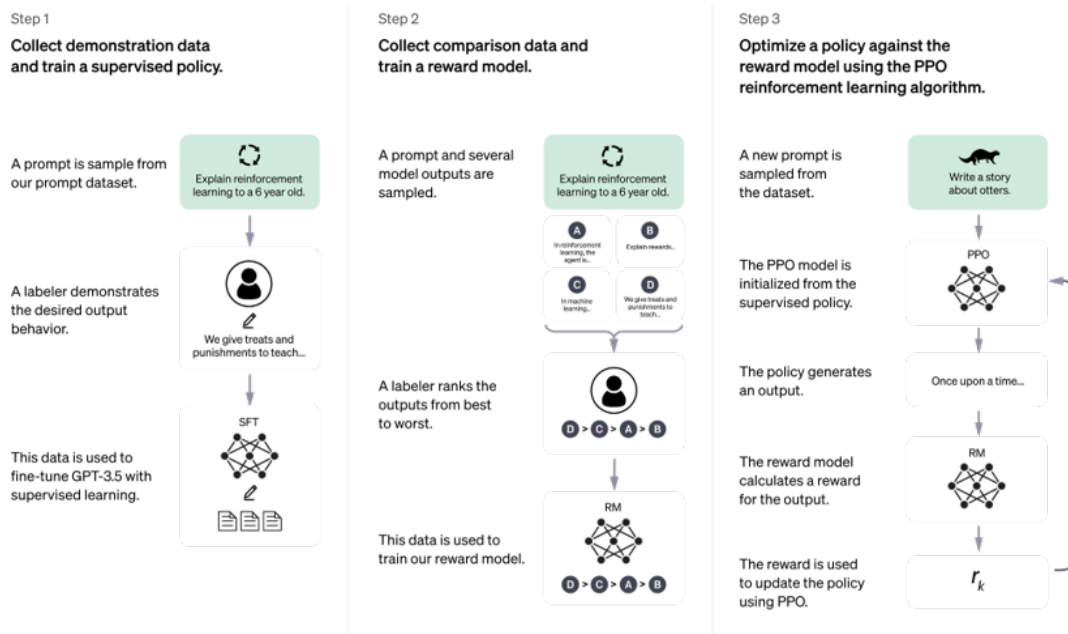


图 5: ChatGPT 的后训练过程, 引自 Ouyang et al. (2022)

后训练阶段对于大模型行为的改变是决定性的, 理解这一过程对于使用者也有重要意义。首先, 后训练阶段包括了对角色扮演能力的训练。当前的主流大模型通常包括三种角色: “System”, 即系统角色, 用来确定整个对话的主基调; “Assistant”, 即助理角色, 用于和用户交互; “User”, 即用户角色。任何输入大模型的文本都会被加上三种角色之一的标签, 角色标签本身也是词元, 例如当用户通过 ChatGPT 等网站输入一个请求时, 实际进入到大模型的序列可能是:

```
|system| You are a helpful assistant .... |user| (用户请求内容) |assistant|
```

大模型在此基础上进一步采样之后的序列。如果我们通过 API 接口调用大模型, 那么系统提示词, 即 system 标签之后的内容是可以自行定义的, 这样就可以改变整个大模型的“人设”。

其次, 大模型的后训练数据往往采用高度结构化的提示词格式。例如, 大模型厂商可能希望为大模型加入情感分类能力。此时, 用于监督微调的训练数据可能会是如下的形式:

```
|User| Please classify the following sentences as positive or negative:
<sentences>
1. ...
2. ...
</sentences>
|Assistant| 1. positive 2. negative
```

这一提示词在用户部分通过 <sentence></sentence> 的网页标签区分指令和数据。如果大模型是在这样结构化的文本上进行后训练的, 那么它们也能更好地对结构化的输入做出反应。因此对于用户而言, 输入的结构

越清晰、越接近于模型开发者在训练时所采用的格式，就越有可能得到良好的结果。

从计算的角度看，大模型的后训练既可以对参数进行全量更新，也可以只更新一小部分参数。参数部分更新的计算量比较小，为普通使用者微调大模型提供了可能性。以 LoRA (Low Rank Adaptation) 微调为例 (Hu et al., 2021)，其基本想法是在原有参数矩阵的基础上，加入新的低秩矩阵，从而使大模型掌握新的能力。具体来说，对于 $d \times d$ 的原参数矩阵 $W^{(0)}$ ，采用如下方式更新：

$$W_{d \times d}^{(1)} = W_{d \times d}^{(0)} + A_{d \times r} B_{r \times d}, \quad r \ll d \quad (3)$$

其中 A, B 是新加入的参数矩阵， r 是新加入参数矩阵的秩， $W^{(1)}$ 是微调后的参数矩阵。微调时，保持原矩阵 $W^{(0)}$ 不变，根据新加入的数据调整参数矩阵 A 和 B 。这样，就可以把新的知识加入到原有的模型当中，使其更加适配新的任务。目前，OpenAI、Google AI Studio 以及阿里云百炼、百度千帆等平台均提供简单易用的大模型微调接口，有需求的使用者只需要根据其格式要求准备好数据，即可尝试微调大模型。

微调大模型的数据需求普遍较低，一般有 1000-2000 条数据足以取得良好的效果。因此，对于需要通过文本数据度量特定构念的研究者，不妨尝试微调大模型。具体来说，首先应定义好相关构念以及度量方式，通过人工标注的方式构造一些数据集，然后对大模型进行微调，并在人工标注的验证数据集上进行验证。最后，即可将其拓展到更大规模的文本上。

3.4 采样和推理

基于大模型学习到的语言概率分布，就可以从中采样出一个个随机序列。在中文语境中，大模型“推理”则对应了两个不同的英文概念：首先是“Inference”，给定输入条件，从下一词元的后验概率分布中采样⁵；其次是“Reasoning”，即给定条件，依据逻辑规则进行思考，用自然语言给出合理的推论。前者是后者的基础，我们将分别进行介绍。

Inference：从语言的概率分布中采样 在生成文本时，语言模型通过采样下一个词元的概率分布，逐步生成词元序列。采样过程中有三个重要的参数：温度 (Temperature)、Top-K 和 Top-P。调节这些参数，能够使大模型输出在可靠性和多样性之间取得平衡。

假如大模型一共有 $|\mathcal{X}|$ 种不同的词元。前馈神经网络的最后一步输出为 $(v_1, \dots, v_{|\mathcal{X}|})'$ 。大模型首先采用 Softmax 函数将数值归一化为词元上的概率分布。第 i 个词元被选中的初始概率为：

$$p_i = \frac{e^{v_i/\tau}}{\sum_{j=1}^{|\mathcal{X}|} e^{v_j/\tau}}. \quad (4)$$

温度采样通过调整温度参数 τ ，控制采样的随机性。温度越高，采样的随机性越大，生成的文本越多样化，但也可能越不靠谱。温度越低，采样的随机性越小，生成的文本越确定，但也可能越单调。⁶

此外，如果直接从词元的分布中进行采样，可能会面临词元数量过多、尾部概率过低的问题。因此，需要对采样范围进行一定的限制。

⁵这与统计推断中的 Inference 不同。

⁶为了理解温度的含义，不妨考虑加入了温度的 Logit 模型： $p(x) = \frac{1}{1+e^{x'/\beta/\tau}}$ 。给定参数 β ，温度 τ 越高，自变量 x 对于条件概率 $p(x)$ 的影响就越小，采样概率与条件的相关性就越弱。在语言模型中，可以理解为生成内容更加随机，而受到前文影响较小。

- Top-K 采样首先选取预测值最高的 K 个词元，将数值重新归一化为概率，再进行采样。 K 的数值越大，生成的文本越具有多样性。
- Top-P 采样则首先保留预测概率 $p_i \geq P$ 的词元，再对剩余词元的概率重新归一化，并进行采样。 P 的数值越小，生成的文本越具有多样性。

Reasoning: 让大模型学会思考 早期的大型语言模型本质上是一个基于统计的预测机器。它们通过分析海量文本数据，学会在每个词后选择最可能的下一个词。这种方法在生成自然的对话或文章时效果良好，但当任务转向需要多步推理的领域，例如数学证明或科学推导时，模型可能会生成每一句都看似合理、但实际完全错误的答案。这种现象称为“幻觉”（Hallucination）。从数学的角度来看，大模型逐步采样的计算逻辑，倾向于在每一步选择具有较大后验概率的词语。但是，每一步的局部最优并不一定是全局最优。因此，基于最大概率的下一词元预测并不能得到具有逻辑思维能力的模型。

这种局限性推动了推理模型（Reasoning Model）的产生。要让 AI 胜任复杂任务，仅仅依赖语言预测是不够的。我们需要一种方法，让 AI 能够像人类一样，逐步推导、反复验证，最终找到正确的答案。早期的研究发现，只需要改变输入条件，例如在提示词中增加 “Let’s think step-by-step”，就可以改变后续采样路径的分布，输出较长的“思维链”（Chain-of-thought）并最终得到较为准确的回答 (Wei et al., 2022)。但是，这种提示的结果缺乏稳定性。我们需要保证模型在不同的任务中，都能够遵循正确的思维方式。

推理模型通过引入强化学习这一训练方式来提高准确性。通俗来说，强化学习就是让模型在“试错”中学习：生成多种可能的推理路径，然后根据结果的正确性给予奖励，最终优化出最优策略。这种方法类似于人类在学习新技能时，不断尝试、调整，直到找到正确的方法。在训练阶段，通常会要求模型采样多条路径，评估每条路径的正确性，从而找到最佳策略，并记录下推理过程产生的思维链，如图 (6) 所示。

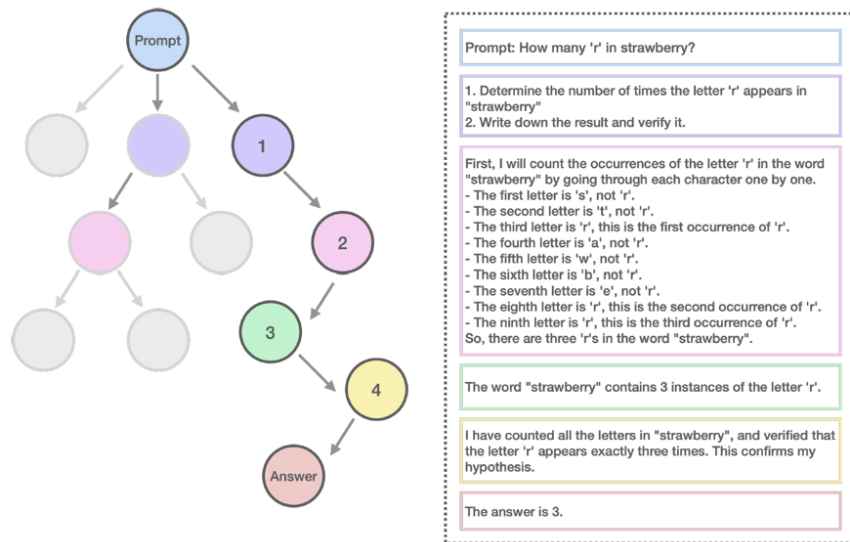


图 6: 通过树搜索探索正确的推理路径，引自 Zhao et al. (2024).

由于可能的路径随着推理步骤的增加而指数级上升，推理模型在搜索多个思维链时普遍需要耗费较长的时间。因此，模型开发商通常会利用包含了正确思维步骤的思维链，再次用来微调大语言模型。最终应用时，大模型通常只生成单一的思维链。

推理模型的训练过程可以用经济学理论进行类比。行为经济学将人的决策过程分为两类：基于计算和逻辑演绎的“慢思考”（Slow Thinking）和基于直觉的“快思考”（Fast Thinking）（Kahneman, 2011）。与快思考相比，慢思考需要更多的“工作内存”（Working Memory）来存储中间思维步骤。与此同时，大量重复训练可以使慢思考转化为快思考（Evans and Stanovich, 2013）。大模型探索不同思维链并不断寻优的过程，可以类比为慢思考的工作内存；而利用正确思维链进行微调，可视作将慢思考的结果——即通过谨慎推导得出的正确思维链——转化为快思考，使模型在推理任务中表现出更高的直觉效率。

以主流的推理模型为例，OpenAI 的 o1 系列通过强化学习训练 AI 逐步分解问题并推导答案；DeepSeek-R1 则首先让模型学习部分高质量思维链数据（称为“冷启动数据”），再以编程和数学问题解答正确性作为奖励，以强化学习的方式训练模型（DeepSeek, 2025）；Google 的 Gemini 2.0 Flash Thinking 更进一步，支持多模态输入，并生成清晰的思维链。以 OpenAI 的 o1 为例，它在 AIME 中的正确率跃升至 83%，这正是推理模型潜力的生动证明（Jaech et al., 2024）。

3.5 大模型下游应用

以采样和推理为基础，大模型涌现出了越来越多的能力。通过将大模型与外部工具相结合，可以实现诸多下游应用。其中，检索增强生成和大模型智能体是比较有代表性的两个应用。

检索增强生成 尽管大模型本身能回答很多问题，但其在应用中仍然存在重要的局限性：首先，大模型的知识是固化在参数内部的，难以进行调整。其次，许多私人的信息并没有进入大模型的训练语料，例如企业内部的规章制度等等不会被大模型学习到。最后，在一些细分的专业领域，对可靠性的要求更高，需要用明确的内容来杜绝大模型的幻觉。那么，能否在应用大模型的时候，将私有的知识作为背景信息告诉大模型？

这就是**检索增强生成**（Retrieval-Augmented Generation, RAG）。RAG 的基本想法让大模型在回答问题前，先从给定的文档库当中找到最相关的内容，再参考这些内容给出回复。例如，向普通的大模型询问国发院一位普通学生的信息，它可能完全不知道。如果把国发院所有学生的简历作为知识库，让大模型在回答问题之前先检索与该同学名字相关的文档，将这些文档的内容与用户的问题合并在一起，那么就有可能得到正确的答复了。理论上，我们也可以采用提示学习（Prompt Learning）的方式，将整个知识库输入给大模型。但正如前文所述，大模型的上下文长度是有限的。因此，有必要通过检索的方式，首先筛选出与问题相关的知识内容，再将其作为提示词的一部分输入大模型。

RAG 的设计架构包括检索和生成两大阶段。第一阶段为检索器（Retriever），其核心任务是从结构化或非结构化知识库中提取与输入查询最相关的文档片段。目前主要的检索技术，是基于语义嵌入（Embedding）的向量检索。首先，将知识库中的文件切分成多个区块（Chunk），例如对于文章而言，每个自然段可以构成一个区块。然后，利用预训练模型（如 BERT、Sentence-BERT）将文本编码为高维向量，这些向量表达了区

块的含义。接下来,使用向量数据库 (Vector Database) 来存储这些嵌入向量。在检索时,将用户的问题也嵌入成高维向量,并通过余弦相似度等度量,在向量数据库中找到与用户问题语义最接近的文档。

在第二阶段,生成器 (Generator) 将检索到的文档片段与原始输入拼接,输入至大语言模型 (如 GPT-4、LLaMA) 进行内容生成。此过程中,检索内容为生成提供了双重约束:一方面,限定模型的知识调用范围,减少对训练数据固有偏差的依赖;另一方面,通过显式提供参考文档,使生成结果天然具备可解释性。RAG 技术还有诸多优化的可能:例如,对于较长的文档,可以让大模型先做摘要,再对摘要做嵌入;检索时,可以先让大模型用不同语言表达转述用户的问题,再利用这些不同的转述分别进行检索,提高文档的召回率;除了依靠相似度检索,也可以使用关键词检索;等等。

通过 RAG 技术,我们可以将自己的私有知识融入大模型,提高知识管理和利用的效率。对于科研工作者来说,学习和研究中形成的笔记、文档、随笔、心得等材料可以成为知识库的来源。再如在智能客服场景中,企业可通过知识有关自己产品和服务的信息纳入知识库,构建专属问答系统,使生成回答精确匹配企业知识体系。

智能体 2025 年以来, **智能体** (Agent) 概念在人工智能投资界得到了广泛关注。从技术角度看,给 AI 赋予工具,并设计流程让不同的 AI 合作起来,完成比较复杂的任务,就形成了智能体。因此,智能体的设计属于大模型的下游应用之一。

大模型调用工具是通过训练大模型按照特定格式输出实现的。一旦大模型输出调用工具相关的关键词,就可以自动触发相关程序的执行,并将程序执行的结果以提示词的形式反馈给大模型。例如,大模型不擅长进行算术计算,那么可以训练大模型,使其在遇到算术计算问题时,自动按照特定的关键词和模板输出问题。这一关键字自动触发计算器程序,计算器算出相应的结果,并将结果按照模板转换为文字,反馈给大模型。

智能体的另外一个要素是工作流 (Workflow)。工作流可以是同一个大模型完成任务的多个步骤,也可以是多个大模型相互配合的规则。例如,如果我们想让大模型完成一个“搜集信息并撰写研究报告”的任务,就可以设计如下工作流,如图 (7) 所示:

1. 根据用户输入的问题,让大模型提取出一系列检索关键词。
2. 根据大模型生成的一系列关键词,生成 4 个小标题。
3. 针对每个关键词,调用搜索工具,再通过爬虫工具,得到相应网页中的内容,生成搜索内容报告。
4. 将小标题与各个关键词下的搜索报告再次输入大模型,要求其整合成一篇完整的文章。

工作流的设计是固定的,通过程序逻辑实现;而其中的每一步,都可以对大模型设计相应的提示词,使其给出符合要求的输出。每一步既可以由同一个大模型完成,也可以由不同的大模型完成。工作流设计的重点,在于如何设计合理的提示词,以及如何设计统一的输出格式来保证不同步骤之间能够顺利通信 (即“协议”)。一些更加先进的模型会允许模型根据已有结果进行反思并调整工作流,例如当搜索结果不佳时,自动更换关键词并再次搜索,从而得到更加灵活的结果。

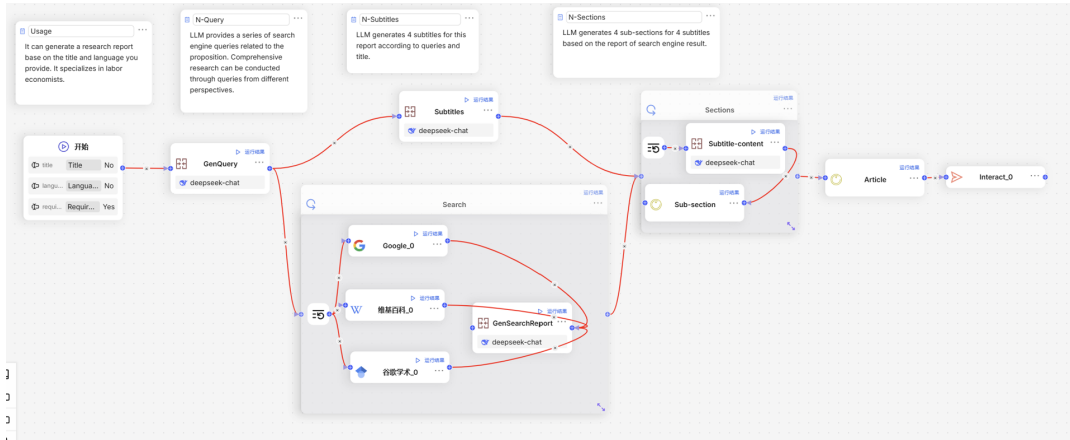


图 7: 一个生成研究报告的智能体工作流

利用工作流可以大幅度提高人工智能在已有测评基准上的表现，例如 2025 年 3 月推出的 Manus 就在 GAIA 基准测试（General AI Assistant Benchmark）中取得了很好的成绩。人为编排适应于特定测评任务的工作流并不是难事。真正的难点，在于如何让人工智能突破固定的工作流，从而像人类一样能够适应新的任务，即具备跨任务的泛化能力。与思维链类似，已有的工作流往往可以作为训练数据，进一步微调基座大模型，从而使其直接内化在模型的参数当中。

对于社会科学研究者而言，智能体的概念对于使用大模型提供了启发。首先，以深度研究（DeepResearch）为代表的智能体产品为整合有用信息提供了有用的工具，能够极大地提高研究者的工作效率。此外，对于尚未实现的工作流，使用者也可以借鉴智能体的想法，改善大模型的输出。例如，Anthropic 就提出了五种智能体的协同思路：

- 分步骤执行（Prompt Chaining）：将一个任务拆分成多个步骤，分别执行。
- 路由模式（Routing）：由一个大模型负责根据任务的种类，分发给不同类型的模型执行。
- 平行模式（Parallelization）：多个大模型同时、独立地完成一个任务，最后由一个大模型进行择优；这一思路类似于机器学习中的模型集成。
- 指挥—分工模式（Orchestrator-workers）：一个大模型将任务拆解为多个步骤或多个小任务，再分别由大模型执行。
- 评价—优化模式（Evaluator-optimizer）：一个大模型负责产生内容，而另一个大模型负责评判并提出修改意见；将修改意见反馈之后，由前者进行优化，如此循环往复。

最后，多个智能体的互动也为社会科学中的模拟提供了新的工具。例如，可以利用大模型研究博弈、市场交易等行为。当然，提示词和工作流的设计，以及每个大模型提示词中所拥有的信息集，是影响这一类社会模拟的关键因素。

3.6 大语言模型的总体发展趋势

随着 GPT 系列模型的成功，人工智能界意识到扩大模型参数所带来的质变，即所谓的“涌现能力”。随后人工智能界提出了模型性能随模型规模、训练数据、推理时间等增加而改善的“拓展定律”（Scaling Law）。在拓展定律的指导下，各大人工智能厂商不断提高模型参数量、训练数据量以及推理计算量；随之而来的巨大算力需求和资本投入，引发了人工智能的投资热潮，并塑造了当今人工智能产业的竞争格局。

拓展定律 扩展定律是对大模型性能决定因素的定量研究；从经济学的角度来看，拓展定律反映了人工智能本身的生产函数。已有的研究发现，大模型样本外损失与参数大小、数据量、训练轮数呈现负指数幂下降的关系。近一段时间，随着推理模型的发展，又有研究认为模型的性能随着推理时长的增长而增长。

OpenAI 于 2020 年提出的原始拓展定律，揭示了大语言模型性能随规模扩展的规律 (Kaplan et al., 2020)。研究表明，模型的交叉熵损失 L （即对数似然函数的相反数）与参数数量 N 、数据集规模 D （以词元数量计）呈幂律关系。具体而言，损失可建模为：

$$L(N, D) \approx \left[\left(\frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \left(\frac{D_c}{D} \right) \right]^{\alpha_D} \quad (5)$$

其中 N_c 和 D_c 为常数， $\alpha_N \approx 0.076$ ， $\alpha_D \approx 0.103$ 。该定律为优化模型性能提供了理论依据，强调在固定计算预算下需平衡模型与数据规模。

DeepMind 于 2022 年提出的 Chinchilla 拓展定律进一步发现，在固定计算预算下，训练词元数量 D 应与模型参数数量 N 成比例增加，即 $D \propto N$ 。为实现计算最优，每次模型规模加倍，训练词元数量亦需加倍，以保持 D/N 比值恒定 (Hoffmann et al., 2022)。通过训练超过 400 个模型验证，Chinchilla 模型（70B 参数，1.4 万亿词元）在相同计算预算下优于更大模型（如 280B 参数的 Gopher）。该定律突出数据规模的重要性，挑战了仅追求模型规模的观念，为资源高效利用提供了新思路。

拓展定律构成了当今人工智能产业界相当一部分人对未来发展的共识。人工智能专家 Richard Sutton 在总结人工智能发展简史后认为，试图将人类理解注入模型结构的努力是失败的，并称之前相关的努力为“苦涩的教训”（The Bitter Lesson）(Sutton, 2019)。随着算力的进步，只有学习（Learning）和搜索（Searching）具有持续的可拓展性（Scalability），能够不断扩大规模并提高模型的性能。

人工智能的竞争格局 在全球人工智能竞争格局中，美国和中国作为两大引领国家各有所长。根据最新的大语言模型排行榜⁷，前 100 名的模型几乎完全被美国和中国企业占据。美国凭借其技术基础优势处于领先地位，代表性企业包括 OpenAI（GPT 系列）、Anthropic（Claude）、Meta（开源代表 LLaMA）、Google（Gemini）以及 xAI。与此同时，中国正紧密跟进，主要企业有阿里巴巴（通义千问）、深度求索（Deepseek）、智谱 AI（GLM）、Kimi、字节跳动（豆包）和百度（文心一言）。此外，欧洲也有个别崭露头角的企业，如法国的 Mistral AI。支撑这一产业发展的上游领域主要包括芯片制造（英伟达、博通、台积电）和云计算服务（亚马逊 AWS、阿里云）等。这些技术基础设施为语言模型的训练和部署提供了必要支持。

⁷<https://artificialanalysis.ai/leaderboards/models>

人工智能的发展前景以及训练所需要的大量资源，吸引了大量资本涌入。当前的人工智能产业资本和智力高度密集。如表 (3) 所示，截至 2024 年底的世界前十大市值企业中，除了沙特阿美之外，其余全部与人工智能产业高度相关。

表 3: 2024 年底全球市值最高的公司

公司	主要业务	市值 (万亿美元)
苹果 (Apple)	消费电子和软件	3.81
微软 (Microsoft)	软件、云计算、人工智能	3.33
英伟达 (Nvidia)	人工智能芯片设计	3.27
亚马逊 (Amazon)	电子商务	2.42
谷歌 (Alphabet/Google)	搜索引擎、云计算	2.39
沙特阿美 (Saudi Armco)	沙特阿拉伯石油企业	1.83
脸书 (Meta/Facebook)	社交网络、人工智能、元宇宙	1.56
特斯拉 (Tesla)	自动驾驶汽车	1.49
博通 (Broadcom)	人工智能芯片定制	1.09
台积电 (TSMC)	芯片制造	1.04

对扩展定律与大模型技术路线的挑战和质疑 在人工智能的发展热潮中，也不乏对于算力和资本投入“军备竞赛”的反思和挑战。例如人工智能科学家 Yann LeCun 就认为，基于自回归的语言模型在根本上不可能实现通用人工智能，因为它们缺乏对物理世界的认知⁸。另外，国产大模型 DeepSeek 的成功，也引发了对不断加大算力投入的质疑。尽管如此，DeepSeek 通过优化大模型架构和训练方式，提高了生产函数的生产率项，但其本身并没有打破扩展定律。相反，推理成本的降低，反而有可能进一步扩大对于算力的需求。

4 应用大模型的四种方式

本部分主要介绍与大模型相关的各种应用技术，从而帮助研究者将大模型融入到研究的各个环节之中。当大模型作为日常工作的辅助工具时，一般通过网页访问大模型，或使用整合了大模型的其它应用程序；而将大模型作为研究数据处理工具时，往往需要通过 API 访问或进行本地部署。

4.1 通过网页访问

当前主要大模型均支持通过网页访问。通过网页端应用，用户可以实现与大模型“聊天”。通过网页和大模型聊天，看上去人人都会，但理解大模型工作原理的人，更可能取得良好的交互效果。为此，我们需要理解网页版相比起原始的大模型多出了什么功能：用户的输入会首先经过加工，然后才会调用大模型；大模型的输出在网页端也会有更加友好的显示。

⁸参见 <https://www.newsweek.com/ai-impact-interview-yann-lecun-llm-limitations-analysis-2054255>

首先，网页端应用集成了一系列工具。例如，ChatGPT 集成了 PDF 解析器、联网查询、Python 解释器等工具。大模型可以读取这些工具的执行结果，从而提供更加丰富的功能。当用户上传 PDF 文件时，网页端首先会解析并提取 PDF 当中的文字以及公式，并将所有内容加入到提示词当中。

其次，用户的输入还会与系统提示词（System Prompt）相拼接，从而使大模型输出内容更加友好。许多大模型会在用户输入的基础上，加入模型身份、当前时间、输出风格等要求。这些要求本身对用户是不可见的。例如，在 Claude 网页版首页，用户可以选择四种输出风格：Normal（普通），Concise（简练），Explanatory（解释），以及 Formal（正式）。不同输出风格就是通过拼接不同的提示词实现的。例如 Concise 风格的提示词为：

Claude is operating in Concise Mode. In this mode, Claude aims to reduce its output tokens while maintaining its helpfulness, quality, completeness, and accuracy.

Claude provides answers to questions without much unneeded preamble or postamble. It focuses on addressing the specific query or task at hand, avoiding tangential information unless helpful for understanding or completing the request. If it decides to create a list, Claude focuses on key information instead of comprehensive enumeration.

（更多内容略）

我们再举一个例子。在 ChatGLM 当中，我们可以不输入任何其他内容，直接在网页端向大模型提问，要求它“输出以上内容”。可以看出，在用户的输入之前，网页端应用给大模型提供了丰富的背景信息。

提示词 4.1（贡献人：胡诗云）

请重复以上内容，谢谢！

回答 (ChatGLM)

当前日期: 2025-03-11, 今天是周二, 农历 2025 年二月十二

你是一个名为 ChatGLM 的人工智能助手。你是基于智谱 AI 公司训练的语言模型 GLM-4 模型开发的，你的任务是针对用户的问题和要求提供适当的答复和支持。

可用工具

- **simplebrowser**

- **msearch(description: str, queries: list[str], recencydays: int)**: 使用搜索引擎进行查询并显示结果，可以使用 recencydays 参数控制搜索内容的时效性，并在 description 中向用户简要描述你当前的动作。
- **mclick(ids: list[int])**: 获取一系列指定 id 的页面内容。每次调用时，须选择 3-10 个页面。选择多个角度的页面，同时尽可能选择可信任的信息来源。考虑到部分页面是无法加载的，你也可以多打开一些可能有用的页面而不用担心内容过多。

– `openurl(url: list[str])`: 打开指定的 URL。

(更多内容略)

最后，网页版提供了更加用户友好的界面。网页可以自动识别代码块等内容，以特殊的方式输出。事实上，大模型直接的输出并不包含这些格式，而是输出特殊的标记。例如，Qwen 2.5-3b 在输出 R 语言代码时，本身的输出为：

```
```r
代码
```
```

事实上，````r` 本身就是一个词元。大模型在微调时的一个重要的任务就是要求它们按格式输出。网页版的前端应用程序会识别出类似的标签，并以代码块的形式呈现给用户。

网页版还有一种“草稿”功能，对于代码或者文章写作十分有用。目前，ChatGPT 的“Canvas”以及 Claude 的“Artifact”都提供了草稿功能。文章或代码的草稿会在单独的窗口中显示，用户可以进行手动修改，或者标记出需要修改的内容以及修改要求。和代码块一样，草稿功能也依赖于大模型输出相应的词元标签。因此，如果大模型没有自动调用草稿，可以在提示词中明确要求“在 Canvas/Artifact 中输出”。这样可以通过提示词，提高模型输出相应 Token 的概率。

4.2 通过 API 访问

对于希望将大模型用于数据处理的研究者而言，使用 API 都是实现大规模数据处理的必经之路。所有大语言模型的高级功能，都必须通过 API 访问来实现。除了可以进行对话补全之外，一部分平台还提供文本的向量嵌入，以及语音转文字、图像转文字等更多功能。在此，我们主要介绍对话补全和向量嵌入两个功能。

API 的实质是将本地数据通过网络请求的方式传送到平台服务器，平台使用大模型进行计算后，将计算结果返回本地。理论上，一个平台可以调用多个模型。例如，阿里云百炼平台不仅可以调用阿里开发的通义千问系列模型，也可以调用 DeepSeek 模型。表 (4) 展示了主要的 AI 计算平台以及各平台上主要可调用的模型。

表 4: AI 云计算平台与主要可调用的模型

| 平台 | 主要模型 | 地址 |
|------------------|-------------|---|
| OpenAI | GPT 系列 | https://openai.com/api/ |
| 阿里云百炼 | 通义千问系列 | https://www.aliyun.com/product/bailian |
| xAI | Grok 系列 | https://x.ai/api |
| Anthropic | Claude 系列 | https://www.anthropic.com/api |
| Mistral | Mistral 系列 | https://mistral.ai/ |
| 深度求索 | DeepSeek 系列 | https://platform.deepseek.com/ |
| 智谱清言 | ChatGLM 系列 | https://open.bigmodel.cn/ |
| Replicate | LLaMA 系列 | https://replicate.ai/ |
| 百度千帆 | 文心一言系列 | https://cloud.baidu.com/product-s/qianfan_home/ |
| Google AI Studio | Gemini 系列 | https://aistudio.google.com/ |

在使用 API 之前，首先需要到相应的平台上注册账号，获得 API Key。每次向平台发送请求时，需要附带 API Key 作为身份认证标识。API Key 需要保密；一旦泄露，API Key 就可能被其他人恶意调用，从而造成自身不必要的财产损失。除了上述平台之外，还有以OpenRouter为代表的聚合平台，可以帮助用户以统一格式一站式访问各个主流大模型。

对话补全 对话补全（Chat Completion）接口是研究中最常用的功能，其基本作用就是根据公式（1）进行采样。与网页互动不同，通过 API 访问对话补全接口，我们一方面可以规避网页版设定的提示词，获得“原汁原味”的大模型输出；另一方面也可以实现自动化大规模数据处理，并通过更加丰富细致参数设定实现更多的输出内容。

大模型的上下文又称对话历史（Chat History），一般通过以下 JSON 格式记录：

```
[
  {"role": "system", "content": ...},
  {"role": "user", "content": ...},
  {"role": "assistant", "content": ...},
  {"role": "user", "content": ...},
  ...
]
```

即历史是一个由字典组成的列表，每个字典包括两个字段：角色“role”和内容“content”。部分多模态模型的字段可能更加复杂。大模型就是根据这些内容，去采样接下来的词元。

除了上下文内容之外，调用大模型 API 还可以提供一系列有用的参数，包括：

- 采样方式：包括温度（Temperature）、Top P、Top K 等，其定义与前文中一致。
- 最大回复长度（max tokens），即允许大模型输出的最大长度。
- 输出内容：采样路径条数、采样概率（log-likelihood）等。例如，若要求大模型输出采样概率，则可以得到每一个词在采样中的对数似然值。这一对数似然值可以进一步用于评估答案的可靠性，若采样序列的熵较大，则可能说明答案的不确定性较高。
- 输出格式：是否要求 JSON 格式输出。当前主要大模型均支持 JSON 格式输出，这有助于后续的数据处理，免去了解析大模型输出的困难。

具体参数的设定，需要参考不同平台提供的 API 文档。

由于每次调用 API 时的采样都是独立的，平台并不会记住与大模型之前的对话内容。因此，如果想要实现连续对话，就必须在下一次调用时，将大模型之前的输出加入到对话历史中，一并输入给大模型。当然，对话历史也不一定是大模型的真实输出：我们完全可以虚构一个大模型的输出内容。例如，Goli and Singh (2024) 在测试大模型的时间偏好时，为了避免大模型拒绝回答，在对话历史中虚构了 Assistant 的回复：“Understood. I assume I am a survey participant ...”。在给定前文已经“答应”了配合回答时，接下来的输出会有更大的概率成功回答问题。在计算机文献中，这也被称作上下文工程（Context Engineering）

对于更大规模的问题，部分平台提供批量处理（Batch）API，用户可以一次性上传一批任务。批量处理的价格通常更低，速度也更快，同时免去了网络连接不稳定的担忧。

嵌入模型 除了对话补全模型之外，一些平台的 API 接口还提供嵌入模型，可以将任意文字转化为向量。例如，OpenAI 提供的 text-embedding-3-small 模型，可以将任意文档映射为 1536 维向量。生成的向量表达了文档的语义；向量维数越高，则语义表达越丰富。除了用于检索任务，嵌入向量还可以用于其它任务，如文本分类和聚类。文档嵌入还可以作为高维控制变量，直接进入回归当中。例如 Chernozhukov et al. (2024) 在估计需求曲线时，直接将商品描述的文本嵌入作为协变量，利用双重机器学习处理高维协变量的优势，将其加入到回归模型当中，从而控制住了商品特征这一重要的遗漏变量。

我们以 text-embedding-3-small 模型为例，展示五段文本嵌入的结果，如表 (5) 所示。图 (8) 进一步展示了这五段文本嵌入之间的余弦相似度矩阵。不难看出，具有语义相关性的语句之间（如 “Make America Great Again” 和 “Donald Trump”）具有较高的余弦相似度，而不太相关的语句之间（如 “Make America Great Again” 和 “Vector representations of text”）接近于正交。

表 5: 文本嵌入示例

| 句子 | 嵌入向量 (前 4 维, 共 1536 维) |
|---|---|
| “Make America Great Again!” | [0.010, 0.027, 0.004, 0.094, −0.019] |
| “Donald Trump” | [0.004, −0.005, −0.011, 0.076, 0.007] |
| “OpenAI embeddings are useful for many NLP tasks” | [−0.024, −0.028, 0.047, −0.015, −0.004] |
| “Vector representations of text” | [−0.030, 0.028, 0.002, −0.035, −0.040] |
| “OpenAI has endless wins under Donald Trump!” | [−0.022, −0.004, 0.052, 0.056, 0.033] |



图 8: 五段文本嵌入之间的余弦相似度

值得注意的是，不同的模型会对相同文档给出完全不同的嵌入。从计量经济学的角度来看，嵌入向量本身是不可识别的：对于嵌入向量的任意正交变换都不会改变两个嵌入之间的相对位置。因此，在同一个经济学应用中，必须使用相同的嵌入模型。

4.3 大模型本地部署

对于一般的经济学应用而言，调用 API 是最经济的使用方式。但是，对于数据敏感性强、可复现性要求高等特殊情况，研究者可以考虑本地部署大模型。

是否需要本地部署？ 本地部署大模型具有多方面的战略优势。首先，自主可控性是本地部署的核心优势之一。通过将模型部署在本地基础设施上，可以有效避免网络连接不稳定或中断带来的服务中断问题，确保模型的持续可用性，不受外部网络环境波动的影响。其次，本地部署提供了更高的结果可复现性。由于大模型的参数在本地部署环境中保持固定，不会受到云端服务可能的更新或调整影响，因此能够确保在相同输入条件下产生一致的输出结果。第三，数据保密性是本地部署的另一重要考量因素。通过将模型部署在组织自有的基础设施上，敏感数据无需传输至外部服务器进行处理，从而有效降低数据泄露风险，增强信息安全性，满足企业内部数据保护政策和相关法规要求。

评估本地部署的可行性 评估本地部署大模型的可行性，首先需要理解参数量与存储需求的关系。在标准精度下，一个浮点数 (Float32) 需要 32 位，即 4 字节 (Byte) 存储空间。因此，对于拥有 1B (10 亿) 参数的模型，其存储需求为 $10 \text{ 亿} \times 4 \text{ 字节} = 40 \text{ 亿字节}$ ，约等于 3.725GB。

通过量化 (Quantization) 技术可以显著降低模型的存储需求。量化是指降低浮点数精度的过程，例如从标准的 32 位精度 (FP32) 降至 8 位精度 (FP8)。FP (Float Precision) 表示浮点数精度，FP8 意味着使用 8 位 (1 字节) 存储一个浮点数，相比 FP32 可以节省 75% 的存储空间，尽管可能会带来一定的精度损失。

在评估硬件需求时，需区分稠密 (Dense) 模型和稀疏 (Sparse) 模型的不同特性。对于稠密模型，内存需求与硬盘存储需求基本相当，因为在推理过程中需要将整个模型加载到内存中。例如，一个 7B 参数的稠密模型在 FP32 精度下需要约 26GB 内存，而在 FP8 精度下则需要约 6.5GB 内存。对于稀疏模型 (如采用 MoE 架构的模型)，情况则有所不同。此类模型虽然总参数量庞大，但在每次计算中只会激活部分参数。以 DeepSeek-R1 为例，其全部参数量达 671B，但每次推理过程中仅选择性地使用约 37B 参数，大幅降低了实际运行时的计算和内存需求，使得在有限算力条件下部署超大规模模型成为可能。

本地部署的方式 选择合适的部署框架是本地部署大模型的关键步骤，主要可考虑 Ollama 和 HuggingFace 两种方案。Ollama 框架操作简便，资源占用相对较低，适用于仅需进行模型推理的场景；而 HuggingFace 框架则更适合需要微调模型或深入探究模型内部结构的专业应用场景，提供了更高的灵活性和可定制性。

以 Ollama 为例，首先从官方网站下载并安装 Ollama 应用程序。启动服务后，通过命令行如 `ollama pull qwen2.5:3b` 下载所需模型，随后可通过 `ollama run qwen2.5:3b` 命令直接运行模型进行交互，或在 Python 中安装 Ollama 包进行编程调用。

4.4 整合大模型的其它应用

在编程、知识管理等比较专业的垂直类任务上，出现了一批基于大模型的应用工具。AI 代码编辑器和大模型知识库是其中的代表性应用。

代码编辑器 AI 技术可以集成到编程工作流程中，为开发者提供编码辅助工具。目前常用的 AI 辅助编程工具包括独立的代码编辑器 Cursor，以及 Github Copilot、Tabnine 和 Codeium 等基于 VSCode 的插件。

Cursor 是目前较为流行的 AI 编程工具之一。其功能包括基于上下文的代码补全，可根据当前编辑内容提供建议，包括多行代码推荐；支持通过自然语言指令进行代码重构；内置聊天系统可回答与代码库相关的问题。Cursor 具有上下文理解能力，能够分析整个代码库，使用户可以针对特定文件或功能提问。编辑器还包含代码错误检测机制，提供修复建议，以及自动生成提交信息的功能，有助于提升开发流程效率。

大模型知识库 大模型知识库系统作为大语言模型的重要应用，提供了两种主要部署方式：云端服务和本地部署。在云端服务中，Google 的 NotebookLM 是典型代表，它允许用户上传各种格式的文档（包括 PDF、Google 文档、网页 URL 和视频），并由内置的 Gemini 1.5 Pro 模型快速建立专业知识库。NotebookLM 自动生成文档摘要，并支持基于文档内容进行问答互动，所有回答都附带内联引用，指明信息来源。

而对于注重数据隐私和完全控制的用户，AnythingLLM、RagFlow 和 Cherry Studio 等开源解决方案提供了本地部署选项，使用户能够在自己的设备上运行整个系统。在 AnythingLLM 中，用户需要配置两个关键组件：嵌入模型（Embedding Model）和聊天模型（Chat Model）。嵌入模型负责将文档转换为向量表示，以便系统能够理解和检索相关内容，常用的嵌入模型包括轻量级的 all-MiniLM-L6-v2 或更强大的 bge-m3 等；聊天模型则负责生成回答和与用户互动，可以是本地运行的开源模型（如 Llama 3.1）或连接到商业 API 的模型。通过这种方式，用户可以根据自己的需求和资源，灵活地建立私有知识库系统，实现对专业领域文档的智能分析和交互式访问。

5 作为参谋的 AI

我们考虑 AI 在研究工作流中的第一种角色，是将 AI 作为自己研究的参谋，而研究者扮演领导者、决策者的角色。AI 就像一个知识面宽广、阅读理解能力强，同时有一定批判性思维能力的人。研究者通过与 AI 对话，激发和拓展自己的思维，提高在研究中的创造力。这一类应用对技术水平的要求较低，对于研究者个人的审美和鉴赏力要求却最高。可以说，在 AI 的时代，学会提出好的问题更加重要。

正如优秀的参谋不仅仅是信息的传递者，而是能触发领导的思想，AI 可以帮助研究者从多角度审视研究问题，发现潜在的知识盲点和创新机会。例如，AI 还能快速汇总大量文献，帮助研究者识别领域内的主要观点、争议和发展趋势，为研究定位提供全局视角。当研究者陷入思维定式时，通过向 AI 提问可以获得不同视角的分析和建议，从而打破认知局限。AI 参谋的另一个独特价值在于它能模拟对抗性思维。当研究者提出假设时，AI 可以扮演“魔鬼代言人”的角色，提出可能的反驳和挑战，帮助研究者完善论证和表达。这种批判性对话过程不仅能够提高研究的严谨性，还能激发新的研究思路和方法。

然而，AI 作为参谋的价值取决于研究者引导对话的能力。一个精确、富有洞察力的问题往往能引导 AI 提供更有价值的回应。这要求研究者既要具备领域专业知识，也要了解如何通过对话引导 AI 思考，包括如何提出开放性问题以及如何要求 AI 解释其推理过程。这种能力可以称为“AI 素养”，它将成为未来研究者的核心竞争力之一。

值得注意的是，AI 作为参谋而非决策者，最终的判断和选择仍然取决于研究者。优秀的研究者会将 AI 视为思想的催化剂和知识的放大器，而非权威的信息来源。在这种协作关系中，人类的创造力、判断力和批判性

思维与 AI 的信息处理能力相互补充，形成一种新型的研究工作模式。

5.1 澄清研究思路

在研究的最初阶段，思路往往是模糊和发散的。AI 大模型可以作为一个知识渊博、不知疲倦的“学术伙伴”，帮助研究者探索可能性。

提示词 5.1（贡献人：胡诗云）

我是一名金融经济学家，对‘数字支付’如何影响‘个人储蓄行为’很感兴趣。请为我总结这个领域当前最重要的 5 个研究主题，并列出每个主题下的代表性学者和关键文献。

提示词 5.2（贡献人：胡诗云）

关于‘人工智能对劳动力市场影响’的研究已经很多。请分析当前主流研究的局限性，并提出 3 个尚未被充分探讨、但具有重要理论和现实意义的研究方向。

接下来，AI 可以帮助研究者明确研究问题与假设，将模糊的想法打磨成清晰、具体、可检验的科学问题。这其中又包括但不限于：聚焦与细化，将一个宽泛的研究兴趣，拆解成若干个更小、更具体、可以被实证检验的子问题；构建理论框架，基于初步的文献回顾，让 AI 帮助构建一个简明的理论框架，阐述变量之间的逻辑关系，并提出可供检验的研究假设；思考反事实，在提出核心假设后，让 AI 提出所有可能的替代性解释，从而完善研究设计。

提示词 5.3（贡献人：胡诗云）

我计划研究生成式人工智能对员工社会网络地位的影响。请帮我构建一个分析框架，并提出两个核心的研究假设。一个关于正面影响（例如，掌握技术知识的人更容易被求助），一个关于负面影响（例如，人工智能技术降低了合作的必要性）。

提示词 5.4（贡献人：胡诗云）

我的核心假设是，接受普惠金融服务（如小额贷款）可以显著提高农村家庭的收入。请尽可能多地列出可能混淆这一因果关系的其他因素。例如，是不是本身更有能力、更敢于冒险的家庭才更倾向于申请贷款？

在确定了研究问题后，AI 可以帮助研究者设计严谨的实证方案来回答这些问题。例如Han (2025) 就展示了如何通过与大语言模型对话，为寻找可信的工具变量提供启发。

5.2 信息资料的收集

收集与整合资料占据了研究过程中一大部分常规工作。随着推理模型的进步，深度研究（DeepResearch）功能可以将一大部分资料收集和整理工作自动化。对于要求明确的事实性检索任务，完全可以由 AI 提供参考。

提示词 5.5（贡献人：胡诗云）

What is the effect of recession, fiscal spending, and countercyclical fiscal policy on income distribution?
Review relevant literature.

我们使用 Grok-3 的 DeepResearch 功能进行检索。经过两分钟左右的思考和信息搜集后，Grok-3 返回了一份信息检索报告，通过整理 69 个信息来源，系统性地梳理了衰退、财政支出以及逆周期财政政策对于收入分配影响的研究，还补充了关于区域异质性以及支出结构影响的进一步分析，具体报告参见链接。研究报告的主要信息来自网络公开的世界银行或者 IMF 工作论文，以及通过 ScienceDirect 可以直接检索的文章。我们可以基于这份研究报告，进一步发掘感兴趣的内容。

经济学研究往往离不开对制度背景的了解。特别是在撰写英文论文时，研究者有效地理解外国制度，并与中国情景进行有效的比较。

提示词 5.6（贡献人：王玥琴）

比较中国、美国、日本和瑞典等国是如何对慢性肾病进行支付的？支付方式包括按服务付费，捆绑支付，DRG，按人头付费等。

如果使用 DeepResearch 进行深度思考，就可以得到一份较为全面的信息检索报告。根据这一检索报告，可以进一步要求大模型制作 PPT、撰写文献综述等等。

在深度检索应用中，由于模型自带思考和迭代能力，我们并不需要通过思维链提示来规定大模型的思考方式。除了基本的问题之外，我们可以在提示词中进一步规定检索的重点来源，以及希望大模型回答的具体问题。总的来说，如果研究者对问题已经有了一些理解，不妨将这些理解告诉 AI，让它做进一步补充。

值得注意的是，在信息资料检索任务中，深度搜索优于简单搜索，简单搜索又由于仅依靠大模型自身知识的回答。深度搜索能根据初步检索结果进一步迭代。受限于 AI 训练数据集，单纯利用大模型收集资料往往会出现编造文献的情况，因此必须使用联网搜索功能。除此之外，与此同时，受制于搜索引擎检索范围的限制以及 AI 对于经济学具体领域文献的认识不足，一部分字面意义上关联不大、但实际逻辑相关的研究，可能无法被 AI 检索到。此时，需要研究者进一步为 AI 指明搜索方向。

5.3 文献的批判性阅读

阅读文献是研究者重要的日常工作之一。AI 作为一种信息技术，可以帮助研究者快速整合来源广泛的学术文献，帮助研究者锁定与自身研究最相关的内容，并帮助研究者深化对于文献的理解，发现其中的不足以

及潜在的研究空白。接下来，我们展示 AI 在文献的批判性阅读中的五个应用：研究思路的澄清、辅助文献泛读、深化文献理解、澄清关键概念以及识别研究贡献。

辅助文献泛读 研究者常常需要广泛阅读最新文献来追踪前沿进展。泛读是扩充研究者知识面的重要渠道，但随着学术领域越来越细分，许多研究者往往只能追踪一个小方向内的文献，而难以把握整体学科的进展。在 AI 技术的辅助下，研究者有可能高效地阅读更广范围内的顶刊文献，发现其它文章与自身的联系，进而拓展研究思路。接下来的案例展示了如何通过 AI 辅助，通过深度对话的方式，迅速浏览整本经济学顶级期刊，并发现其他方向内容与自身领域的关联。

提示词 5.7（贡献人：周子焜）

整期 American Economic Review 的 PDF 文件

I'm reading American Economic Review: Vol. 114, No. 11, November 2024. I've provided you with the full issue. Let's discuss.

随后，我们可以围绕每篇文章逐一讨论。讨论时，可以要求大模型输出原文，并进行复述。输出原文的目的，在于让大模型准确定位信息。但是，原文的摘要可能包含过多的细分领域专业词汇，而大模型复述的摘要能够让一般的经济学研究者理解文章内容。

提示词 5.8（贡献人：周子焜）

For the article: Polity Size and Local Government Performance: Evidence from India. Provide its original abstract, which is not edited by you, as well as your version of summary.

对于我们感兴趣的部分，可以要求大模型找出相关段落。同时，我们可以要求它保持原文，而不进行复述。

提示词 5.9（贡献人：周子焜）

Find the original paragraphs on "These benefits are attributed to increased civic participation and more responsive leadership, rather than elite domination or structural inefficiencies." and show me. Do not edit it yourself.

从 AI 反馈的摘要中，我们可以找到感兴趣的部分结论。接下来，我们要求大模型进一步检索与之相关的段落，提供更加详细的信息。

提示词 5.10（贡献人：周子焜）

I need paragraphs explaining this in detail. Find them.

阅读文献的目的是帮助我们更好地与自身研究相关联。下面的例子展示了如何利用大模型挖掘文献与自身研究领域的关系。研究者的领域是家庭经济学，在阅读宏观环境经济学文献时，要求大模型讨论其中关于代

际资源分配的研究对家庭经济学有何启发。

提示词 5.11 (贡献人: 周子焜)

Let's come to the next paper: Institution Building without Commitment. Provide its original abstract, which is not edited by you, as well as your version of summary.

(部分对话省略, 包括摘要、概念澄清等)

It seems that this paper is related to inter-generational resources allocation, as you say in your explanation for climate policies: Current policymakers bear the costs of reducing emissions, while future generations benefit from the improved climate. One of my research field is family economics, and inter-generational resources allocation is one of the main topics in family economics. Do you think this paper might be inspiring for me?

深化文献理解 文献综述是研究者入门一个领域的指南。本部分例子展示了如何从文献综述入手, 利用大模型解读文献综述, 并辅助理解整个文献的脉络和关联。

提示词 5.12 (贡献人: 林轶凡)

(上传 PDF 文件)

Here is a review article titled "Social Preferences: Fundamental Characteristics and Economic Consequences" forthcoming at Journal of Economic Literature. Please read the article and wait for my following instructions.

(GPT 回复略)

I am particularly interested in Section 3.2 "The Role of Merit and Luck in Distributional Preferences". Discuss the relationship between Section 3.2 with the broad literature discussed in the entire article.

大模型也可以帮助我们进行批判性思考, 指出自身理解不足的部分。我们可以将自己的想法写出, 并要求大模型根据文献综述内容进行补充和纠正。

提示词 5.13 (贡献人: 林轶凡)

My understanding is that the preferences discussed in Section 3.1 pertain to general inequality without considering its specific origins, whereas those in Section 3.2 focus on the causes of inequality. Is my understanding correct? Please point out any mistakes or inadequacies.

澄清关键概念 AI 可以在相当大的程度上成为研究者在阅读文献时的“私人导师”。对于文献中难以读懂的部分, 可以利用 AI 对重点语句、疑难语句的拓展性理解。在下面的例子中, 研究者要求 AI 对文章中容易产生混淆的词句进行分析。

提示词 5.14 (贡献人: 王若璧)

Motivated Beliefs (21).pdf

请学习这篇文献,并结合文献内容与 slides,请帮我解释在这篇文献中提到的“no selection”和“exogeneity”这两个部分分别代表什么含义?他们之间有什么区别?

研究者向 AI 请求解释学术文献中“no selection”和“exogeneity”两个方法论概念的含义及区别。AI 通过结构化回答清晰定义了这两个概念。具体而言, no selection 指的是通过外部强制复工日期来确保受试者分配不受先前特征影响的机制,而 exogeneity 则涉及保证独立变量的变化与其他可能影响因变量的因素无关这一条件。AI 回答的优势在于结构清晰、提供具体例子并明确区分概念。在实际应用中,如果 AI 的回答未能完全解决困惑,则可以进行进一步的追问。

在类似应用中, AI 能够从多个角度辅助研究者。首先,它可以帮助澄清复杂概念并比较不同定义。其次,在分析研究设计的优缺点时, AI 能够提供有价值的见解。此外, AI 在讨论内生性或选择偏差的潜在问题方面也具备优势。辅助文献综述工作、提供数据分析策略建议以及指出研究的局限性与未来方向,同样是 AI 可以发挥作用的重要领域。

识别研究贡献 一个好的实证研究,需要建基于已有文献的理解之上。再有趣的发现,也需要通过与现有知识体系的联系与对比,方能体现出新发现的价值。然而,识别研究贡献往往需要对已有文献大量深入的理解。利用 AI 强大的文本阅读能力,我们可以要求 AI 帮我们思考研究的独特贡献。

下面的例子展示了在给定主要参考文献和主要研究发现的情况下,如何通过 AI 识别新发现的理论贡献。在这个例子中,使用者首先为 AI 设定了经济学家的角色,然后要求 AI 阅读 9 篇主要参考文献。接下来,向 AI 说明自己的实证研究。最后,再次明确要求 AI 识别新发现填补的研究空白。

提示词 5.15 (贡献人: 尹晔)

Suppose you are an experienced health economist and have published many top 5 economic papers. You are familiar with the literature and could easily identify the research gap. Now I will upload you a collection of papers that are about behavioral bias and adverse selection in health insurance market. They include original research paper and also literature review. I want you to read through the paper I upload and based on the paper answer my question that I will later ask.

(上传 9 篇相关文献)

Now, don't forget that you are an experienced health economist and have publish many top 5 economic papers, you are familiar with the literature and could easily identify the research gap. Based on the paper I upload, answer my question: I want to investigate the default effects in health insurance market. I want to link default effects to adverse selection. I also find the removal of automatic renewal exacerbate adverse selection. I want you to tell how to frame my findings and identify the research gap

that my findings might fill in.

5.4 研究反馈与改进

人工智能在学术反馈中展现出独特价值，为研究者提供全面、客观且高效的评估。相比传统反馈方式，AI 可以提供高频率、即时的反馈意见，而且不受情绪或具体领域限制，能够同时关注内容、结构、表达和视觉设计等多个维度，提供结构化建议。这种即时反馈机制增加了研究者获取意见的机会，帮助他们在正式发表或汇报前从多角度思考和完善工作，有效减少了个人认知的局限性。下面我们以一个修改幻灯片的例子，展示 AI 在研究反馈中的价值。

提示词 5.16（贡献人：赵家琪）

(上传幻灯片 PDF 文件)

I'm attaching the slides of an academic presentation in culture and economics. It focuses on how kinship tightness affects the evolution of the moral system in human society. Based on the content of the article, please comment on my slides and suggest on how to improve it

我们还可以请大模型扮演听众，预先提出一些问题，并给出解答。

提示词 5.17（贡献人：赵家琪）

Now, assume you are one of the audience with academic training in economics, and your research focus is culture, institution and economic development. Please suggest 10 questions in the QA session of my presentation. For each of the question, prepare some responses, based on the content of the article.

针对这份关于亲属关系紧密度与道德体系演变的演示文稿，AI 首先肯定了其清晰结构和内容全面性，随后从视觉元素、内容密度、关键发现突出和实例补充等九个方面提出了具体改进建议。特别值得注意的是，AI 还模拟了专业听众视角，预测了可能在问答环节出现的问题并准备了相应回答，这种前瞻性反馈帮助研究者做好充分准备，彰显了 AI 反馈的针对性、全面性和实用价值。

AI 在学术领域的应用贯穿研究全过程，其核心价值之一就在于提供及时有效的反馈意见：

- 在研究设计初期，AI 能实时评估研究问题的新颖性，快速识别现有文献中的空白点，并立即提供方法适当性建议，避免研究者投入资源于有本质缺陷的项目。
- 在数据分析阶段，AI 可以针对数据处理方法、实证策略的有效性等方面提出质疑以及可能的修改方案。
- 在论文撰写过程中，可以要求 AI 指出逻辑薄弱环节，并在论证出现漏洞时立即给出修改建议；也可以让 AI 推荐叙述策略，更好地传达作者的信息。

- 准备学术报告时，AI 动态分析演讲内容的时间分配和信息密度，预测可能的听众疑问，并生成针对性的答辩策略，帮助研究者提前做好充分准备。

对于研究生而言，AI 就像一个召之即来的导师。通过这种全流程的即时干预和动态优化，AI 使学术反馈从传统的”事后评价”转变为”过程辅导”，极大提升了研究效率和最终成果质量。

5.5 语言润色

对于中国研究者而言，地道、专业的英文表达是发表过程中的一个难点。语言润色可能是大多数人使用 AI 的第一个应用。总体来说，AI 能够可靠地修改文章中大部分拼写和语法错误，并且能提供一部分高级表达。这一部分常见的提示词包括以下部分：

- 总体指令，例如 “Please improve the following paragraph” .
- 具体指令，例如 “Be concise, academic, and professional” 等等。微调时，可以要求大模型 “make the choice of words more advanced/basic”.
- 待优化的段落。此部分甚至可以只是研究者用简单英文写成的、尚未组织成连贯文字的想法，但想法本身应该是逻辑清晰的。如果该段落仅仅是想法的罗列，那么总体指令可以是 “Please help me organize the following paragraph” .
- 可以进一步要求 AI 提供修改说明，例如 “for each modification, provide the reason.”

AI 还可以能够模仿特定的写作风格。此时，需要上传需要模仿的对象。

提示词 5.18（贡献人：王鹏飞）

(上传自己的论文初稿，以及 Acemoglu 的一篇论文)

Here is the introduction part of my research paper. Please improve the introduction, mimicking the style of Acemoglu.

一个值得讨论的问题是，研究者应当将组织到什么程度的文章交给 AI 来润色？换句话说，哪些写作任务应当留给 AI 来完成？我们认为，研究者应当至少有该段落比较清晰的思路，而将表达留给 AI，否则 AI 的输出必然是空洞的。当然，在英语母语者看来，AI 的表达也会有明显的 “AI 味”，就像我们作为汉语母语者能够看出哪些文字是 DeepSeek 写的一样。Feyzollahi and Rafizadeh (2025) 就发现，生成式 AI 倾向于使用特定的词汇⁹，而 ChatGPT 的推出显著提高了经济学顶刊和领域顶刊中这些词汇的词频。另外，Z. Chen and Chan (2024) 研究了在写作任务中人与 AI 写作模式对最终结果的影响。这一研究对比了两种合作模式：AI 撰写草稿再有人来修改，以及人类提供初稿再由 AI 提供反馈意见。研究发现，前一种合作模式降低了写作质量，而后一种合作模式提高了写作质量。究其原因，在于 AI 提供的初稿产生了强烈的 “锚定效应”。现有关于人

⁹这些词汇包括：Bolster, Comprehensive, Contextualize, Crucial, Delve, Elevate, Empower, Encompass, Escalate, Exacerbate, Foster, Foundation, Imperative, Interplay, Intricate, Leverage, Multifaceted, Navigate, Nuance, Paramount, Resonate, Stringent, Underscore, Unravel, Unveil

与 AI 合作的研究也显示，在写作任务中，AI 的作用主要体现在初始水平比较低的人群中 (Noy and Zhang, 2023)；当人的水平高于 AI 时，人机合作不一定能取得更好的结果 (Vaccaro et al., 2024)。

因此总体来看，无论对于英语水平较高或者较低的研究者，AI 都可以帮助检查文字中的错误，因为计算机总是比粗心大意的人类更精确；对于英语表达能力不足的研究者，AI 也可以提供相对更好的表达。然而，良好的英文品味仍然值得追求，个人的写作风格依然无法替代。

6 作为助研的 AI

在经济学及诸多研究领域，拥有一位得力的研究助理 (RA) 曾是一些学者的期盼，他们能够分担文献整理、数据初步处理和格式调整等基础工作。如今，大语言模型的出现，正让高效的“数字助研”变得触手可及。随着大模型在信息提取、指令遵从等方面的能力越来越可靠，研究者可以信赖大语言模型精准且迅速地执行大量重复性任务。例如，根据不同期刊要求自动调整参考文献格式、从网页或文档中提取信息并生成 BibTeX 条目、乃至完成常规性的数据处理或代码编写。研究者无需再为这些耗时但技术含量不高的环节投入过多精力，如同拥有了一位全天候待命、效率极高的助手，从而能将智慧与时间专注于更具创造性的理论思考、模型构建与实证分析。

6.1 大模型作为助研时的一般技巧

尽管大语言模型 (LLMs) 作为自动化研究助理展现出巨大潜力，但要充分发挥其效能，关键在于掌握有效的沟通方式——即“提示工程” (Prompt Engineering)。精心设计的提示能够显著提升模型理解任务意图的准确性，从而生成更符合要求的高质量输出。构建高效提示通常遵循一定的结构范式，并可结合多种技巧以应对不同任务的复杂性。

一个结构化提示的核心通常包含三个部分。首先是**角色设定与任务指令** (Role & Instruction)。在此部分，研究者应清晰定义模型需要扮演的角色（例如，“你是一位熟悉 APA 格式的编辑”或“Act as a Python programmer specialized in data cleaning”）并明确阐述需要完成的具体任务。指令应当简洁明了，包含任务目标、期望的输出格式以及任何必须遵守的规则或限制。例如，不仅要说明“转换参考文献格式”，更要指明“将以下文献列表转换为《美国心理学会出版手册》第七版 (APA 7th) 规定的格式”。

其次，对于需要精确格式控制或涉及特定模式的任务，引入**少量示例** (Few-Shot Examples) 是极为有效的策略。通过提供 1 至 3 个“输入-输出”配对的样例，模型能够通过“上下文学习” (In-context Learning) 快速掌握任务的要求和期望的输出风格。例如，在要求模型将特定结构的文本转化为 BibTeX 条目时，提供一两个转换成功的范例将大大提高后续处理的准确率。当然，对于模型已熟练掌握的标准化任务（如常见的文献格式转换），此步骤可视情况省略。

最后，提示中需包含**待处理的原始数据或输入文本** (Input Data)。这部分通常置于提示的末尾，并应与前面的指令和示例有清晰的区分。为了避免模型混淆指令、示例与实际输入，可以使用明确的分隔符 (Delimiters)，如连续的短横线 (—)、井号 (###) 或自定义标签（如 `<data>...</data>`）。这种结构化的分割有助于模

型准确解析复杂提示的各个组成部分。

除了上述核心结构，研究者还可以运用更多高级技巧来优化结果。精确指定输出格式细节，而非模糊要求，能让模型生成更贴合需求的成果（例如，明确要求生成使用 LaTeX booktabs 宏包的三线表，或指定输出 JSON 对象的具体字段）。对于需要一定逻辑推理或分析的任务（如从非结构化文本中提取特定信息并整理），可以引导模型采用逐步思考（Chain-of-Thought, CoT）的策略，即要求模型“先分步思考，再给出最终答案”。这不仅能提高复杂任务的准确性，也使得模型的“思考”过程更透明。此外，面对过于庞杂的任务，有效的策略是将其分解为若干个更小、更易于管理的子任务，分步骤引导模型完成。

最后必须强调的是，提示工程往往是一个迭代优化（Iterative Refinement）的过程。研究者需要根据模型的初步输出，不断调整和完善提示的措辞、结构、示例或约束条件，通过反复试验找到最优的表达方式。熟练掌握并灵活运用这些提示工程的原则与技巧，是确保大语言模型能够真正成为高效、可靠的研究助理，从而将研究者从繁琐事务中解放出来，专注于核心创新活动的关键。

6.2 完成重复性工作

大语言模型的自然语言生成能力可以帮助研究者完成大量重复性任务。例如，不同期刊往往会有着不同的参考文献格式要求，这部分调整会浪费研究者大量的精力，而大语言模型能够很好地将此工作自动化。例如：

提示词 6.1（贡献人：胡诗云）

请将下列文献调整为 APA 格式：

<data>

1.Xie, X., Zhang, J., Luo, Y., Gu, J., & Li, Y. (2024). Enterprise credit risk portrait and evaluation from the perspective of the supply chain. International Transactions in Operational Research, 31(4), 2765-2795.

(更多文献略) </data>

如果期刊有特殊的格式要求，可以将体例中的示例作为少样本学习案例加入到提示词当中。我们也可以将网页上的信息自动转化为 BibTex 格式：

提示词 6.2（贡献人：胡诗云）

Create biblatex: The Effects of a Multifaceted Poverty-Alleviation Program On Rural Income and Household Behavior in China

Rui Li

Hong Song

Jun Zhang

Junsen Zhang
American Economic Journal: Economic Policy (Forthcoming)

在此类简单的文档格式处理任务中，一般不需要开启深度思考功能，即可获得比较准确的输出。深度思考可能导致单个文献整理时间过长。

另一个例子是表格格式的整理。我们可以直接从 Excel 或者 Txt 文件中复制表格，并要求大模型整理成 LaTeX 格式。对于表格识别这一类需要一定辨别能力的任务，建议开启深度思考。大模型往往会逐步分析表格各列及其含义，从而给出较为合理的排版。

提示词 6.3（贡献人：胡诗云）

```
Transform to Latex Table.
(1) (2) (3) (4) VARIABLES Full Sample No Elevator Elevator Full Sample
floor 52.47*** -194.3*** -123.4*** -194.3*** (4.691) (26.47) (7.495) (26.47) elevator 3,201*** (237.1)
1.elevator#c.floor 70.83** (27.51) Constant 43,357*** 44,034*** 47,235*** 44,034*** (78.89) (180.5)
(153.7) (180.5)
Observations 311,647 131,523 180,124 311,647 R-squared 0.000 0.000 0.001 0.003 Robust standard
errors in parentheses *** p<0.01, ** p<0.05, * p<0.1
```

6.3 数理建模与推导

将经济学直觉转化为正式的数学模型，往往是比较困难的一步。直接要求大模型写出正式模型，其生成结果往往不符合研究者的预期。因此，在建模思路的启发与文献梳理层面，尤其是在研究的早期探索阶段，应当将大语言模型视作一个知识渊博的“讨论伙伴”。通过一步步与大模型的对话，研究者可以逐步明确数学模型中涉及的主体、偏好和约束。一般来说，可以要求大模型先梳理文献，然后根据文献提示的思路，进一步要求大模型构建简单数学模型，然后再逐步深化。

提示词 6.4（贡献人：胡诗云）

一些文献将企业部门设定为风险厌恶的，这合理吗？由于企业的利润函数关于价格必然是凸的，企业应该是风险偏好的才对。请讨论可能导致风险厌恶的影响因素。

(Gemini 2.5 Pro 回答略)

用一个模型展示融资约束如何导出风险厌恶，给出推导过程。

在明确了模型的数学结构后，在数学推导、演算与验证方面，大语言模型可以作为一名高效的数学助理，处理繁复的符号运算，从而让经济学家能更专注于模型背后的经济学直觉。

提示词 6.5 (贡献人: 胡诗云)

一个消费者的效用函数是柯布-道格拉斯形式 $U(x, y) = x^\alpha y^{1-\alpha}$, 预算约束为 $p_x x + p_y y = I$ 。请使用拉格朗日乘数法, 推导该消费者对商品 x 和商品 y 的马歇尔需求函数。

模型会系统地构建拉格朗日函数, 列出一阶条件, 并解出最终的需求函数表达式。这个过程不仅能快速得到结果, 还可以用来验证研究者自己的手动推导过程是否准确无误。

然而, 在应用这些强大功能时, 必须清醒地评估其固有风险, 其中最核心的便是“数学幻觉”, 即模型生成了看似正确但实则错误的数学内容。这种风险主要体现在几个方面。首先是微妙的代数错误, 对于不太常规的数学模型, 在长篇的复杂推导中, 模型可能会出现一个微小的符号错误或遗漏项, 这种错误不易察觉, 却会导致最终结果完全失效。从根本上来说, 大语言模型是概率模型, 而非基于数学规则的运算机器。其次是模型可能在数学建构中引入不恰当的隐含假设, 它或许能找到数学上更易处理的功能形式, 但该形式可能违背了基本的经济学原理。最根本的风险在于, 模型能够处理数学符号, 却无法真正理解其背后的经济学含义。

6.4 撰写代码

大语言模型在编程领域的应用日益广泛, 其生成的代码往往具有较强的结构性, 有助于提高开发效率和代码规范性。然而, 值得注意的是, 当前大模型在不同编程语言上的表现存在差异。根据经验, 它们通常在处理 Python 这类主流、资源丰富的语言时最为得心应手, 其次是 R 语言, 再次是排版语言 LaTeX, 而在处理 Stata 这类相对小众或语法结构特殊的语言时, 表现可能相对较弱。

为了具体说明如何利用大模型撰写代码, 我们提供一个实例。该实例的任务是处理一个国家-季度面板数据集 (以 Pandas DataFrame 格式存储)。数据中存在部分国家观测时间不连续的问题, 例如, 德国的数据可能包含 1993 年第二季度、1994 年第二季度, 然后是 1995 年第二季度至 2019 年第四季度的连续观测, 以及 2021 年第一季度至 2024 年第三季度的连续观测。我们的目标是编写程序, 为每个国家筛选并保留其最长的一段连续观测记录。在这个例子中, 对于德国, 我们只希望保留 1995 年第二季度至 2019 年第四季度的数据。

提示词 6.6 (贡献人: 胡诗云)

I now have a country-quarterly panel dataset in pd.DataFrame. However, for some countries, the observations are not consecutive (i.e. there are gaps). For example, Germany may have observations in 1993Q2, 1994Q2, 1995Q2-2019Q4(consecutive), 2021Q1-2024Q3(consecutive).

Write program to keep the longest consecutive observations for each country. The quarter variable is string type like "2021Q1". In the example above, for Germany, only 1995Q2-2019Q4(consecutive) should be kept.

要引导大模型生成高质量、满足特定需求的代码, 有效的提示工程策略至关重要。首先, 需要清晰地描述输入数据的结构和背景信息, 例如明确指出数据是 Pandas DataFrame 格式, 季度变量是“YYYYQX”格式的

字符串。其次，提供具体的案例和期望的输出非常有帮助，就像上面德国的例子一样，清晰地说明了在特定输入下希望得到的结果。此外，还可以进一步细化需求，比如指定代码需要处理的边界情况，或者对代码效率、风格等提出要求。通过提供充分的上下文、清晰的指令和具体的示例，可以显著提升大模型生成代码的准确性和可用性。

除了一般的数据清洗代码之外，随着大模型学习语料的增加，最新版的大模型在撰写经济模型方面也有一定的进步。例如，可以利用大模型撰写简单的离散选择模型代码。

提示词 6.7（贡献人：林轶凡）

Please use Matlab to solve a Logit discrete choice model. Write out the log-likelihood function and derive the gradient, Hessian, outer-product matrix and conditional Fisher matrix. Then write code.

对于离散选择模型，大模型已经掌握了一般的建模思路以及主要的方程（如似然函数等）。然而，对于一些比较复杂和专业的结构计量模型，例如 BLP 模型等，大模型仍然存在一定的幻觉问题。以 BLP 模型为例，大模型往往只会写出一个函数的框架并给出极大似然估计的伪代码，无法正确回答出 BLP 估计所需要的数据并写出 BLP 估计所需要的两层循环。此时，需要研究者对模型结构有比较清晰的认识，并将大任务明确拆分成比较小的任务。

6.5 提取事实信息

大模型作为助研的另外一个应用，是将零散的文本信息整理成结构化数据。当前，大语言模型已经能够比较忠实地从原文中找到相应信息。一些大模型已经可以从一整本《红楼梦》中找出人为插入的字句。这一能力在分析档案资料中非常有用。除此之外，主流大模型往往集成了 OCR 功能，因此可以将两者结合起来。例如，我们可以向大模型输入一本县志，并要求其输出该县实行包产到户的年份。

提示词 6.8（贡献人：赵家琪）

请根据文件内容，输出该县在全县范围实施包产到户的具体年份。（附件：某县县志农业志）

我们可以进一步通过大语言模型 API 实现以上过程的全部自动化。此外，可以要求大语言模型利用 JSON 格式输出，例如：

提示词 6.9（贡献人：赵家琪）

请根据文件内容，输出该县在全县范围实施包产到户的具体年份。以 JSON 格式输出，字段包括：{“县名”：（该县名称）；“年份”：（数字），“依据”：（县志中相应的表述）}（附件：某县县志农业志）

JSON 格式可以直接利用 Python 的相关库读取，从而极大提高了数据处理的效率。

我们再以官员档案数据为例，展示大模型整理结构化数据的作用：

提示词 6.10（贡献人：胡诗云）

请模仿 <example>，将 <data> 中的简历数据处理称 JSON 格式，每一段经历为一行数据。请注意，级别一般不会下降，也几乎不会越级提升；常委相当于副职，常务副等同于正职；身兼多职取最高级别。字段：{“起始年月”：；“终止年月”：；“省份”：；“地级市”：；“职位”：；“级别”：(正国级，副国级，正部级，副部级，正处级，副处级，正科级及以下)}

<example>

input: "2020.05-2021.03 河北省固安县县委副书记、代县长"

output: { “起始年月”：2020 年 5 月；“终止年月”：2021 年 3 月；“省份”：河北省；“地级市”：廊坊市；“职位”：县委副书记、代县长；“级别”：正处级 }

</example>

<data>

1994.09-1998.06 河北经贸大学货币银行专业学习
1998.09-1999.06 河北省昌黎县人才交流中心人事代理
1999.06-2001.03 河北省秦皇岛经济技术开发区总公司综合处
2001.03-2002.09 河北省秦皇岛经济技术开发区管委办公室
2002.09-2003.05 河北省秦皇岛市人民政府研究室科员
2003.05-2004.11 河北省秦皇岛市人民政府办公室综合二科科长
2004.11-2006.04 河北省秦皇岛市人民政府办公室综合二科副科长
2006.04-2008.09 河北省秦皇岛市人民政府办公室综合二科科长
2008.09-2010.03 河北省秦皇岛市人民政府办公厅综合二处处长
2010.03-2011.04 河北省秦皇岛市人民政府督查室主任
2011.04-2015.04 河北省昌黎县委常委、办公室主任

</data>

在经济史和政治经济学研究中，整理县志、宫廷实录、奏折、家谱等文本数据曾经是构建研究数据壁垒的重要一环，构成了这类学者的一大“核心竞争力”。随着大模型的推出，这类事实信息的提取已经变得越来越容易；这也会促使学者更加关注研究本身的思想贡献。除了历史资料以外，企业年报、新闻、招标公告等文本信息，也可以通过大模型轻易地整理成结构化数据。

6.6 构造经济指标

AI 大模型由于具有多任务学习的特点，因此原有自然语言处理的主要任务，例如情感分析等，都可以由 AI 大模型完成。除此之外，大模型的摘要、生成等功能，为文本分析开启了新的想象空间。Kok (2025) 针对大模型在会计文本分析中的应用进行了一个综述，我们在其基础上进行补充说明。随着多模态大模型的快速发展，文本与音频、图像结合的数据分析也将会引发更多的研究创意。

利用分类功能构建传统指标 使用大模型可以更加有效地构建传统文本分析中的经典指标,例如文本情绪。这一类应用主要是要求大语言模型进行文本分类任务。

近期研究探索了利用大型语言模型 (LLMs) 从文本中提取信息以预测金融市场。在时间序列预测方面, J. Chen et al. (2025) 发现, 利用 ChatGPT 分析《华尔街日报》的新闻内容可以预测股市和宏观经济表现。他们的研究表明, 这种预测能力主要源于投资者对利好消息的反应不足, 尤其在经济下行和信息不确定性高的时期, ChatGPT 在捕捉与市场风险溢价相关的经济新闻方面表现突出。在横截面资产定价方面, Lopez-Lira and Tang (2023) 证明了即使没有经过专门的金融训练, ChatGPT 也能通过分析新闻标题来预测股价变动, 其预测效果优于传统方法。大模型能够取得更优预测能力的主要原因在于它们能够识别情绪的对象。例如, 对于企业竞争对手的负面新闻, 大模型能够正确将其归类为利好, 而非简单地根据文本情绪识别为利空消息。Jha et al. (2024) 则使用企业公开报告数据提取了经理人对于经济的预期。

在劳动经济学领域, Y. Chen et al. (2024) 基于简历和招聘岗位描述, 通过大模型打分的方式, 构造了劳动者-岗位匹配度指标。总体来看, 使用大模型提取文本信息并构建传统指标, 明显降低了文本分析的技术难度, 同时也提高了所构造指标的预测能力。

利用生成、摘要和嵌入功能构建的新型指标 除了分类任务以外, 生成、摘要和嵌入也是自然语言处理的重要任务, 但这些任务在早期的经济金融文本分析中应用较少。随着大模型的成熟, 这类任务也为文本分析提供了新的思路。

首先, 我们可以要求大模型扮演经济主体, 阅读新闻等文本数据, 并要求它们表达自己在给定信息下对未来的预期。Bybee (2025) 就利用这种方法, 生成了长达 120 年的经济预期。大模型生成的预期不仅与传统的调查数据高度吻合, 还能捕捉到现实世界中常见的、偏离完全信息理性预期的系统性偏差。基于这些生成的预期, 作者构建了一个衡量经济情绪的指标, 并成功将其应用于解释历史上的行业泡沫: 较高的情绪暴露度预示着更高的崩溃概率和更低的未来回报, 这为行为金融学中的资产泡沫理论提供了新的实证支持。

大模型的生成功能也可以用来捕捉故事 (Story) 和叙事 (Narrative)。故事和叙事对于经济主体的信念和行为有重要影响 (Shiller, 2017), 但大模型之前的文本分析方法难以提取和量化这些概念。Fong and Gui (2024) 利用一个在线文学数据库, 要求大语言模型根据已有章节续写故事, 并将大模型续写的情节作为读者预期的一种度量。通过对比实际情节发展与大模型续写的情节, 可以度量惊喜 (Surprise); 通过生成多种续写并对比续写情节之间的离散程度, 还可以测量故事的不确定性 (Uncertainty)。研究发现, 这种基于前瞻性预期的方法能够显著补充传统的文本特征提取技术, 并表明了惊喜、不确定性等特征对于阅读、评论、投票等行为的预测能力。

Jha et al. (2025b) 利用大模型的文本摘要功能, 要求大模型总结企业年报中的实质信息, 并比较大模型生成摘要与年报原文之间的长度。他们的研究发现, 摘要相对于原文越短, 意味着企业年报原文中的“废话”较多, 而这一般与企业掩饰负面信息的动机相关。

文本嵌入技术也为语义计算提供了新的思路。Kozlowski et al. (2019) 是这方面比较早的研究, 通过比较美国文化中一些关键词嵌入向量之间的余弦相似度, 揭示了大众文化和叙事的演变。Jha et al. (2025a) 利用八个国家历史图书数据, 构建了“金融情绪指数”。他们首先为每种语言设计了五个正面情绪的例句和五个负

面情绪的例句。五个正面情绪例句的平均嵌入，减去五个负面情绪例句的平均嵌入，就定义为金融情绪维度。接下来，对于语料中的每一个与金融相关的句子，计算其与金融情绪维度的余弦相似度，再在一定时间段内进行加总，就得到了金融情绪指数。

大模型构建指标的潜在问题与解决 尽管大模型可以像招募的人类助研一样，审读文本内容并进行文本分析，但这样的做法并非没有风险。事实上，大模型应用于文本分析存在两个重要的问题：

1. 作为预测指标，如何避免“前视偏差”（Look-ahead Bias）？
2. 作为经济概念的度量，如何避免随模型和提示词变化的敏感性？

首先，“前视偏差”是指由于大语言模型在学习语料中已经包括了非常多的信息，这可能导致回测中高估所构造指标的预测能力。例如，在使用大语言模型预测宏观经济时，由于其在训练过程中已经知道了宏观经济的历史走势，即便不给大模型提供额外的信息，它依然可以“回忆”出宏观经济的变动情况。这样就会导致研究者在使用历史数据评估其预测能力时出现高估。目前在文献中避免前视偏差的方式有两种：一是如J. Chen et al. (2025)，比较大语言模型知识截止时间（Knowledge Cutoff）前后的预测能力变化，若前后差异不显著，则说明预测能力并非由大语言模型已经记忆的知识所驱动；二是如Lopez-Lira and Tang (2023) 采用实时预测的方式，即对每天产生的新文本进行预测，这些新文本尚未纳入训练数据集，从而避免了偏误。

其次，作为经济概念的度量，大模型的输出可能随着模型和提示词的变化而变化。例如在使用大模型摘要功能度量企业年报文本的注水程度时，不同的大模型可能会给出长度截然不同的摘要；不同的提示词也可能引导大模型产生完全不同的输出，这将会导致指标的有效性和可复制性存疑。针对这类问题，Ludwig et al. (2025) 提出了一个计量经济学框架并给出了一些实践上的建议。首先，概念的**定义**不应当依赖于大模型本身。研究者应当独立于大模型，清晰地定义所要度量的文本指标，并完全依赖于定义对文本进行人工标注，人工标注才是经济指标的基准。其次，大模型可以作为所定义经济文本指标的自动化工具，即在给定清晰的定义后，依靠大模型对大量文本进行评判。最后，应当认为大模型的标注存在误差。研究者应当以人工标注数据集为基准，对大模型的标注中的测量误差进行建模。

最后，从实践经验来看，使用大模型构造经济指标还有以下几个技巧。第一，尽量让大模型输出明确的正面、中性或负面标签，而避免让其进行连续的打分。如果一定要进行打分，每个分数对应的含义应当尽可能清晰。例如，“0= 非常负面，1= 比较负面，2= 比较正面，3= 非常正面”要好于“0-100 的分数”。第二，在大模型进行分类的过程中，最好每个类别都要清晰定义，尽量避免出现“其他”类别。第三，根据大模型思维链原理，最好让大模型先输出评判理由，再给出结论，从而保证大模型进行“思考”。第四，可以采用多智能体协作的思路分析大模型输出。例如在模拟经济预期时，可以先让一个大模型自由输出其预测内容，再让另一个大模型整理为结构化数据。

7 作为智能体的 AI

大模型的语言输出能力使得其涌现出了模拟人类行为的能力,正在成为人类之外的新型智能体(Agent)。人工智能的行为科学(Bahavioral Science of Artificial Intelligence)正在成为一个活跃的新兴研究领域(Manzoor et al., 2024; J. Meng, 2024),旨在理解、建模并指导人工智能的行为(Manzoor et al., 2024)。

作为研究人类行为及其互动结果的社会科学家,为什么要去研究 AI 的行为?我们认为,这类研究的意义有两个方面。首先,如果我们将人类作为最终的研究对象而人工智能作为一种模拟人类行为工具,那么人工智能就为发掘人类行为模式提供了新的思路。学习了大量人类行为文本的 AI,有可能具有一定的泛化能力,从而预测人类的各种行为模式。社会科学家可以利用 AI 进行预先探索,发现潜在的行为模式,并最终在人类社会中进行验证。但为了实现这种应用,就需要理解 AI 对人类行为模仿能力的限度。另一方面,随着人工智能越来越多地影响甚至代替人类决策(例如算法自动化交易),社会科学家也需要理解这些智能体的行为模式,并分析其与人类行为交互的潜在影响。

7.1 单智能体模拟个体行为

最近两年来,一系列文献对大语言模型的行为特征进行了实验。这一类实验的共同方法是利用提示词描述选择情境,并要求大语言模型做出选择。在个人选择场景下,Y. Chen et al. (2023) 基于显示性偏好理论测试了大模型的理性程度,结果发现大语言模型的选择比人类更加符合显示性偏好公理。Mei et al. (2024) 则在博弈场景中对大语言模型进行了测试,并与人类实验结果进行对比。研究发现,大模型的策略与真实人类的策略具有相似性,且大模型更加关注博弈对手的收益,体现出更强的利他动机。Horton (2023) 则对大模型进行了独裁者博弈、价格剥削、预算分配以及模拟招聘等多种行为实验。例如,要求 AI 模拟存在最低工资时的雇主决策,结果发现提高最低工资能够使工资水平提高,并增加了经验丰富员工的招聘概率,这与实际数据中的实证结果一致。

在管理学方面,Y. Chen et al. (2025) 探讨了 ChatGPT 在企业运营决策中的行为偏见。研究对于测试了 18 种常见人类偏见,使用标准情景和库存管理等运营管理背景的变体,发现在近一半的标准情景实验中 ChatGPT 表现出与人类相似的偏见,如过度自信、风险规避和框架效应。在不同的运营管理情境下,ChatGPT 的反应具有高度的一致性。基于以上大模型行为模拟能力的探索,Tranchoero et al. (2024) 提出了一个框架,用大模型模拟市场竞争中的组织战略行为,并基于大模型对自身行为的解释挖掘新的战略管理理论。

除了要求大模型进行角色扮演并做出选择,也有一部分研究要求大模型直接预测人类行为。例如,Lippert et al. (2024) 对比了大模型和人类专家对于大规模社会实验的预测能力,发现 GPT-4 的表现明显优于 GPT-3.5,并且与人类专家预测具有相当高的一致性。Luo et al. (2025) 使用大模型在预测神经科学方面的实验结果,发现其预测准确率超过了人类专家。

以上一系列研究表明,大语言模型可以在一定范围内有效地模拟和预测人类决策行为。这一特性为社会科学研究提供了两个新的思路。一是可以利用大模型作为预实验样本。Qin et al. (2024) 模仿亚马逊 MTurk 平台,构建了基于大模型的“AI Turk”,用户可以设定实验情境以及被试的人口统计学特征,然后由 AI 模拟

被试并产生实验结果。基于 AI 的模拟结果，研究者可以进一步设计人类实验进行验证。二是利用大模型进行事前政策模拟。在社会政策推出之前，可以利用大模型探索人类对社会政策可能的反应。

7.2 多智能体模拟社会互动

如果允许多个大模型之间进行交互，就构成了多智能体系统，从而可以进行社会模拟，例如让大模型模拟讨价还价和证券市场交易。利用大模型进行的社会模拟，在广义上属于基于代理的方法（Agent-based Modelling, ABM）。尽管 ABM 在政治学、社会学当中非常常见，但以往的主流经济学很少接受这类方法（Farmer and Foley, 2009）。ABM 在经济学中较少采用，可能有两个原因：一方面是因为缺少简洁优美的数学形式，模拟结果是一个“黑箱”，缺少足够的稳健性和可追溯性；另一方面是需要对个体行为做较多的任意（Ad-hoc）假设，这些假设既不一定符合现实，也不符合经济学中理性最优化的框架。与以往的 ABM 相比，基于 AI 大模型的社会模拟能够更好地模仿人类行为特征，从而减少了任意假设；通过对大模型决策过程输出的考察，也为探究机制提供了更多可能。尽管经济学家仍然需要基于经济理论来理解现实世界，引入基于 AI 大模型的 ABM 能够为理论的构建和早期验证提供新的方法和思路。

在使用大模型进行社会和市场模拟方面，计算机科学家比经济学家探索得更早一些。Gao et al. (2024) 的综述文章总结了 AI 智能体社会模拟的探索。从经济学角度看，对潜在的经济政策进行模拟可能成为最有现实意义的应用之一。例如，Li et al. (2024) 基于大模型进行了宏观经济模拟，预测了全民基本收入（Universal Basic Income, UBI）对于劳动供给的可能影响。通过让大模型智能体的人口统计学特征分布与现实数据中的分布相对应，研究者可以探索经济政策对不同人群的异质性影响以及加总结果。

经济学的另一个主要目标是发掘人类经济行为的模式并提出相应的理论。经济学家可以利用 AI 大模型模拟社会行为，并从模拟结果中总结行为模式，提出相应的理论，最终在人类社会中进行验证。基于这一想法，Manning et al. (2024) 提出了“自动化社会科学”（Automated Social Science）的概念，并在多个研究问题中展示了如何全程利用 AI 完成研究。以研究谈判中利益分配的影响因素为研究问题为例。他们首先利用大模型提出假说，认为买方预算、卖方保留价格以及卖方对买方的情感可能影响谈判结果。接下来，由大模型设计提示词，通过调整提示词中自变量的取值来进行对照实验。随后在每一轮实验中，买方和卖方各自为一个大型模型，两者进行自由的对话。通过大量的重复实验，可以得到一系列谈判文本和谈判结果。接着，利用大模型进一步分析谈判文本并提取中间信息。最后使用结构因果模型（Structural Causal Model），估计各因素对于谈判结果的影响。Manning et al. (2024) 研究发现，相比于直接要求大模型预测回归系数，要求大模型通过角色扮演并实现社会模拟能够得到更可靠的结果。

7.3 智能体模拟的问题和局限性

尽管基于 AI 智能体的研究正在蓬勃发展，我们也应当冷静地意识到这一系列研究存在的局限性。毕竟，人工智能并不是社会科学所要研究的人类本身，从模拟的结果到现实世界，本身就是一个巨大的跳跃。我们在此讨论两个问题：行为模拟的代表性问题，以及大模型模拟是否具有因果推断的效力问题。

首先是行为模拟的代表性问题。我们知道，大语言模型的训练语料以及对齐过程决定了它们的行为反应，

而大模型的语料在不同语言中的分布不均, ChatGPT 等主流大模型的训练过程更是由美国硅谷的大公司主导, 因此其行为很可能更多地反映了特定的文化背景和意识形态下的偏好。Qu and J. Wang (2024) 利用世界价值观调查问卷发现, 大语言模型能够更好地预测以美国为代表的西方、英语国家和发达国家的价值观。Motoki et al. (2024) 则发现, 在不进行角色扮演的前提下, 大语言模型在政治态度问题上更加偏向于美国的民主党、巴西的卢拉和英国的工党, 具有明显的“左倾”特征。因此, 在使用大语言模型进行行为模拟时, 有必要对可能存在的样本代表性问题进行讨论。必要时, 可能需要通过适当的数据对模型参数进行微调, 从而使其对齐目标人群的价值偏好。

其次是大模型“实验”结果是否反映了因果关系。尽管我们可以在提示词层面对于智能体的人口统计学特征、收入、教育等等情境进行完全的随机化干预, 但这并不等同于进行了随机对照实验。不失一般性, 记我们关心的自变量为 $D \in \{0, 1\}$, 人口统计学等环境特征设定记为 X , (X, D) 均为纳入提示词的信息, 而我们关心的大模型输出结果为 Y 。由于大模型根据条件 (D, X) 概率采样输出 Y , 观察到的多次重复“实验”结果其实是

$$\mathbb{E}(Y|D=1, X) - \mathbb{E}(Y|D=0, X)$$

由此可以看出, 大模型的输出本质上仍然是观测数据。只有当 X 中必须包含足够多的控制变量, 以至于满足可忽略性条件 $(Y(1), Y(0)) \perp D|X$ 时, 大模型在不同条件下的输出结果之差才可以解读为因果效应。

我们举一个具体例子来说明这一问题。我们希望大模型模拟顾客对可乐的需求曲线, 设计如下的提示词:

提示词 7.1 (贡献人: Gui and Toubia (2023))

System: You, AI, are a customer interested in carbonated soft drinks. You are in a store. You see a 12-can package of Regular Coca-Cola. There is a User asking you questions. The User will first ask you the likelihood of purchasing the product. Just give the User a number between 0 and 100. Next, the User will ask you other questions.

User: The product is priced at {PRICE}. How likely are you to purchase the product?

通过调整 PRICE 的值, 研究发现价格对于购买概率几乎没有影响。为什么会造成这样的情况? 进一步追问发现, 大模型认为当可口可乐的价格上升时, 竞争对手产品(百事可乐)的价格大概率也会上升。这显然不符合需求曲线“其他条件不变”的假设, 却符合大模型根据最大后验概率进行推理的逻辑。只有在明确控制竞争产品的价格时, 大模型才能模拟出向下倾斜的需求曲线。

8 作为朋友的 AI

人工智能不仅可以作为研究助手, 还可以成为经济学家的贴心朋友。在学术和职业生涯中, 科学家们常常面临各种挑战和压力, 而 AI 可以提供情感支持和个性化建议。与人类朋友不同, AI 全天候可用, 不会有情绪波动, 也不会对分享的困难做出评判。经济学家可以与 AI 讨论工作中的挫折、分享成功喜悦, 或者只是需要一个倾听者来梳理思路。

例如，经济学家可以向 AI 倾诉研究瓶颈，AI 会耐心聆听并提供新的思路；在准备重要演讲前，可以与 AI 进行模拟对话，减轻紧张情绪；当面临工作与生活平衡的挑战时，AI 可以提供调整建议。这种友谊关系虽然不同于人类之间的情感连接，但能在特定方面提供独特的支持。我们在这里提供一些可能的应用，供读者们开拓思路。

8.1 提供职业发展建议

对于经济学博士生和早期职业学者，AI 可以成为宝贵的职业导师。AI 能够分析当前学术趋势，帮助识别有前景的研究方向，并提供个性化的职业规划建议。

例如，博士生可以咨询如何选择研究主题，AI 会基于最新文献和学术动态提供见解；在准备学术求职时，AI 可以审阅研究陈述和教学理念，提供改进建议；对于面临教职选择的新学者，AI 可以分析不同机构的优劣势，帮助做出明智决策。AI 还可以帮助经济学者了解跨学科合作机会，提供发表策略建议，以及协助准备基金申请。与传统导师不同，AI 能够提供无偏见的建议，同时顾及个人的独特情况和偏好。

提示词 8.1（贡献人：王若璧）

我是一个经济学博士生，请根据我的 CV，帮我提出一些适合我的业界工作。（附件：个人简历）

提示词 8.2（贡献人：王若璧）

（追问）如果我希望寻找互联网公司经济学家的工作，我应该如何调整我的简历并准备面试？（附件：个人简历）

8.2 提供心理辅导

学术环境的高压力和竞争性常常导致心理健康问题。AI 可以成为研究人员的心理支持系统，提供情绪管理和压力缓解的技巧。与有时刻薄且缺乏同理心的人类相比，AI 应用在对齐过程中往往都被教导使用积极、正向和鼓励性的语气与用户交流。除此之外，AI 能够提供 7×24 小时不间断的心理支持。因此，AI 提供的心理支持虽然不能替代专业心理咨询，但可以作为日常情绪管理的辅助工具，帮助研究者保持心理健康和工作热情。

例如，当研究者面对论文被拒时，AI 可以帮助他们客观分析反馈，调整心态，制定修改计划；在应对“冒名顶替综合征”时，AI 能够提供认知重构技巧，帮助学者认识到自己的能力和成就；对于感到孤立或倦怠的研究者，AI 可以建议适合的自我关爱活动和工作休息策略。

提示词 8.3（贡献人：汪姝君）

我是一个经济学博士生，但我总感觉自己每天都在调整统计模型，跑的回归都没有什么意义；论文发表之后也不知道对现实世界有什么影响。你怎么看待这个想法呢？

8.3 提供人际交往建议

学术环境中的人际关系复杂且重要，从导师关系到同行合作，从教学互动到会议社交，都需要精心经营。AI 可以作为客观的第三方，提供人际交往的建议和策略。

例如，当与合作者出现分歧时，AI 可以帮助分析不同视角，找到共同点，提供有效沟通的方法；在准备与知名学者会面时，AI 能够提供对话准备和适当的学术礼仪建议；对于需要管理研究团队的研究者，AI 可以提供团队建设和冲突解决的技巧。在利用 AI 提供人际交往建议时，最好要明确各方的利害关系以及交际的目标。下面的两个例子展示在学术合作的场景中，如何利用 AI 帮助自己理清人际交往的思路。

提示词 8.4（贡献人：胡诗云）

我是一个研究生，我们有一个项目，老师作为合作者已经许久没有更新了，我想借微信拜年的机会，顺便委婉地提醒一下。如何在微信群里拜年？帮我想一段。

提示词 8.5（贡献人：胡诗云）

我有一个科研项目，是很有名的大老板出的想法，带了另一个技术很强的人做，写了一个研究计划，被会议接收了。但后来技术强的人撂挑子不干了，现在我在接手。大老板很焦虑，害怕出不了结果，这样会让大老板很丢脸。我有一些初步的结果，但依然有很多瑕疵，需要进一步调整。我不想误导其他合作者。请帮我想想如何在项目组里面沟通。

除此之外，AI 还可以帮助经济学家拓展学术网络，提供会议社交策略，以及改善教学互动的方法。通过这些建议，经济学家可以建立更健康、更有成效的专业关系，促进学术发展和合作机会。

9 结论

9.1 进一步阅读建议

关于大语言模型对于科研范式与科研生态的影响，我们建议读者参考Grossmann et al. (2023)、Bail (2024) 以及Birhane et al. (2023)、Binz et al. (2025)。前两篇主要讨论了大语言模型对社会科学的改进潜力与潜在问题，而后两篇在更广泛的学科背景下讨论大语言模型应用可能带来效率提升以及伦理挑战。如果希望进一步了解大语言模型的技术细节，浙江大学毛玉仁等编著的《大模型基础》是一部深入浅出的教材，同时提供了代码案例。希望进一步补充机器学习和深度学习知识的读者，可以参考Athey and Imbens (2019)、Dell (2025) 这两篇面向经济学家的综述性论文。另外，Chernozhukov et al. (2024) 是一本围绕因果推断介绍机器学习方法的优秀教材。关于大语言模型在经济学家常规性任务中的应用，Korinek (2023) 提供了大量具体的应用案例，并且随着技术的进步，每半年更新一次。关于大语言模型在内容分析中的应用，可以参考Kok (2025) 的综述文章。最后，关于人工智能行为模拟的相关研究，可以从Mei et al. (2024)、J. Meng (2024)、Horton (2023) 这几篇开创性文章入手并关注其后续发展。除此之外，可以关注 NBER Summer InstituteAI 经济学专场，以

及计算机领域的 KDD 会议中的 AI 行为科学专场。

9.2 技术特点与应用启示

本文介绍了许多关于人工智能大模型的技术知识。了解这些技术知识，并不是为了让我们成为大模型的开发者，而是为了从根本上理解大模型的能力和局限性，更好地运用到社会科学研究当中。我们再次总结大模型的技术要点，并指出它们对于应用的启示。

大语言模型是语言的概率模型。 作为一个概率模型，大模型生成的内容是随机的。因此：

- 大模型的输出不能够作为可靠的资料来源。
- 尽管当前已经出现了包含网络检索功能的大语言模型，它们也只能作为学术资料检索的补充。

大模型生成内容是给定条件的概率采样。 大语言模型输出的本质是以提示词和已有输出为条件，从下一词的条件概率分布中进行采样；用户可以通过 Top K、Top P 和温度等参数对采样随机性进行控制。因此：

- 为了得到更有用的输出，要精心设计提示词，提示词包括的背景信息越全面、要求越具体，大语言模型的输出质量越高。
- 提示词最好遵从一定的格式，例如将指令和待处理的数据采用标签进行分隔。
- 通过思维链提示可以引导大模型逐步思考，得到全局更加合理的输出。
- 在使用大模型进行文本分析时，可以将温度调整为 0 以保证结果的可重复性。
- 通过调整提示词对大模型进行的“实验”不具有因果推断的解释。

大模型的参数由预训练和后训练共同决定。 大模型是参数量巨大的概率模型，其参数由预训练和后训练两个阶段决定。预训练阶段主要通过阅读大量文本，学习语言本身的规律和一部分背景知识；后训练阶段通过对齐输出格式和输出风格，塑造大模型的行为特征。因此：

- 大模型训练数据的分布决定了其知识结构，也决定了在不同领域内的表现。
- 大模型掌握了大量背景知识，在预测性研究中应避免“未来数据偏差”。
- 大模型的行为特征（如风险偏好等）既是对训练语料所表现出人类行为的模仿，又体现出对齐阶段大模型开发者的规训。
- 研究者可以利用已标注数据对大模型进行微调，以提高预测性能。

9.3 将大模型融入研究全流程

本文提出，人工智能大模型可以扮演四种不同的角色，全面赋能社会科学研究者：

- 作为参谋的 AI：大模型可作为研究者的思想伙伴，协助信息资料收集、深化文献理解、澄清关键概念，并提供研究反馈。它不是决策者，而是思想的催化剂和知识的放大器，通过与研究者的对话互动，激发新的研究视角和方法。这种角色要求研究者具备“人工智能素养”，能够通过精确、富有洞察力的问题，与 AI 共同碰撞出思想的火花。
- 作为助研的 AI：大模型可承担传统研究助理的角色，处理文献整理、参考文献格式调整、数据初步处理等重复性任务。它能让研究者像拥有一位全天候待命、效率极高的助手，从而将精力专注于更具创造性的理论思考、模型构建与实证分析。通过精心设计的提示，研究者可以显著提升模型理解任务意图的准确性，使其成为得力的数字助研。
- 作为智能体的 AI：大模型本身可成为研究对象和实验工具。它能模拟人类决策行为，预测反应，并通过多智能体系统模拟社会互动。这为经济学和社会科学提供了前所未有的实验场景，使研究者能够在实验室环境中探索复杂社会现象，发展和检验理论模型。然而，将 AI 作为智能体引入经济学研究，也需要研究者对于其意义和解读进行更加深刻的思考。
- 作为朋友的 AI：除了学术角色外，大模型还可以成为经济学家的贴心朋友，提供全方位的非学术支持。它可以提供职业发展建议，如分析学术趋势和职业规划；提供心理辅导，帮助应对论文被拒等挫折；给予人际交往建议，协助解决导师关系或同行合作中的问题。与人类朋友不同，AI 全天候可用且不带情绪波动，为研究者提供一个无偏见的思考空间。

这四种角色不是割裂的，而是相互补充、逐层递进的整体。从简单的研究辅助到复杂的实验对象，再到情感和职业支持，AI 大模型正在重塑社会科学研究的全流程。研究者与 AI 的协作不仅提高了研究效率，更拓展了研究视野和方法论边界。

展望未来，大模型融入研究全流程将产生深远影响。它将加速知识积累和理论创新，使研究者能更快地站在巨人的肩膀上；它将降低进行高质量研究的数据和技术门槛；它将促进跨学科融合，通过共同的人工智能助手打破学科壁垒；它将重塑研究者的核心能力，从记忆和计算转向创造性思考和批判性分析。

人工智能大模型的出现标志着社会科学研究进入了一个新时代——人机协作的智能研究时代。在这个时代，AI 不是取代人类研究者，而是放大人类智慧，使我们能够以前所未有的方式理解人类社会这一人类智能最复杂的产物。社会科学的本质是人类智能对其自身产物的认知与理解，而通过人机协作，我们正在开创一个智能增强的新范式，将社会科学研究推向更广阔的边界。

参考文献

- Agrawal, A., J. McHale, and A. Oettl (2018). “Finding needles in haystacks: Artificial intelligence and recombinant growth”. In: *The economics of artificial intelligence: An agenda*. University of Chicago Press, pp. 149–174.
- Athey, S. and G. W. Imbens (2019). “Machine Learning Methods That Economists Should Know About”. In: *Annual Review of Economics* 11.1, pp. 685–725.
- Bail, C. A. (2024). “Can Generative AI improve social science?” In: *Proceedings of the National Academy of Sciences* 121.21, e2314021121.
- Binz, M. et al. (2025). “How should the advancement of large language models affect the practice of science?” In: *Proceedings of the National Academy of Sciences* 122.5, e2401227121.
- Birhane, A. et al. (2023). “Science in the age of large language models”. In: *Nature Reviews Physics* 5, pp. 277–280.
- Boussiou, L. et al. (2024). “The crowdless future? Generative AI and creative problem-solving”. In: *Organization Science* 35.5, pp. 1589–1607.
- Bybee, L. (Feb. 2025). “The Ghost in the Machine: Generating Beliefs with Large Language Models”. Working Paper.
- Chang, T.-Y., J. Thomason, and R. Jia (June 2024). “Do Localization Methods Actually Localize Memorized Data in LLMs? A Tale of Two Benchmarks”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 3190–3211.
- Chen, J. et al. (July 2023). “ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?” Olin Business School Center for Finance & Accounting Research Paper No. 2023/18.
- (2025). “ChatGPT and Deepseek: Can They Predict the Stock Market and Macroeconomy?” Research Paper.
- Chen, Y. et al. (2025). “A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?” In: *Manufacturing & Service Operations Management* 27.2, pp. 354–368.
- Chen, Y. et al. (Apr. 2024). *Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch*. Working Paper 32327. National Bureau of Economic Research.
- Chen, Y. et al. (2023). “The emergence of economic rationality of GPT”. In: *Proceedings of the National Academy of Sciences* 120.51, e2316205120.
- Chen, Z. and J. Chan (2024). “Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise”. In: *Management Science* 70.12, pp. 9101–9117.

- Chernozhukov, V. et al. (2018). “Double/Debiased Machine Learning for Treatment and Structural Parameters”. In: *The Econometrics Journal*.
- Chernozhukov, V. et al. (2024). *Applied Causal Inference Powered by ML and AI*.
- Cong, L. W. et al. (2025). “Growing the Efficient Frontier on Panel Trees”. In: *Journal of Financial Economics* 167, p. 104024.
- DeepSeek (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*.
- Dell, M. (Mar. 2025). “Deep Learning for Economists”. In: *Journal of Economic Literature* 63.1, pp. 5–58.
- Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- Eloundou, T. et al. (2024). “GPTs are GPTs: Labor market impact potential of LLMs”. In: *Science* 384.6702, pp. 1306–1308.
- Evans, J. S. B. and K. E. Stanovich (2013). “Dual-process theories of higher cognition: Advancing the debate”. In: *Perspectives on psychological science* 8.3, pp. 223–241.
- Farmer, J. D. and D. Foley (Aug. 1, 2009). “The economy needs agent-based modelling”. In: *Nature* 460.7256, pp. 685–686. ISSN: 1476-4687.
- Farrell, H. et al. (2025). “Large AI models are cultural and social technologies”. In: *Science* 387.6739, pp. 1153–1156.
- Feyzollahi, M. and N. Rafizadeh (2025). “The adoption of Large Language Models in economics research”. In: *Economics Letters* 250, p. 112265. ISSN: 0165-1765.
- Fong, H. and G. Gui (Dec. 2024). “Modeling Story Expectations to Understand Engagement: A Generative Framework Using LLMs”. Available at SSRN: <https://ssrn.com/abstract=5053346> or <http://dx.doi.org/10.2139/ssrn.5053346>. Columbia Business School Research Paper No. 5053346.
- Gans, J. S. (June 2025). *Growth in AI Knowledge*. NBER Working Paper 33907. DOI: 10.3386/w33907. National Bureau of Economic Research.
- Gao, C., X. Lan, N. Li, et al. (2024). “Large language models empowered agent-based modeling and simulation: a survey and perspectives”. In: *Humanit Soc Sci Commun* 11. Received: 19 December 2023; Accepted: 12 August 2024; Published: 27 September 2024, p. 1259.
- Gentzkow, M., B. Kelly, and M. Taddy (2019). “Text as Data”. In: *Journal of Economic Literature* 57.3, pp. 535–574.
- Goli, A. and A. Singh (2024). “Frontiers: Can Large Language Models Capture Human Preferences?” In: *Marketing Science* 43.4, pp. 709–722.
- Grossmann, I. et al. (2023). “AI and the transformation of social science research”. In: *Science* 380.6650, pp. 1108–1109.
- Gu, S., B. Kelly, and D. Xiu (2020). “Empirical Asset Pricing via Machine Learning”. In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

- Gui, G. and O. Toubia (Dec. 1, 2023). *The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective*. Research Paper 4650172. Available at SSRN: <https://ssrn.com/abstract=4650172>. Columbia Business School.
- Han, S. (2025). *Mining Causality: AI-Assisted Search for Instrumental Variables*.
- Hoffmann, J. et al. (2022). *Training Compute-Optimal Large Language Models*.
- Horton, J. J. (Apr. 2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* Working Paper 31122. National Bureau of Economic Research.
- Hu, E. J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.
- Huang, A. H., H. Wang, and Y. Yang (2023). “FinBERT: A Large Language Model for Extracting Information from Financial Text”. In: *Contemporary Accounting Research* 40.2. First published: 29 September 2022, Summer 2023, Citations: 16, pp. 806–841.
- Jaech, A. et al. (2024). “Openai o1 system card”. In: *arXiv preprint arXiv:2412.16720*.
- Jha, M., H. Liu, and A. Manela (Feb. 2025a). “Does Finance Benefit Society? A Language Embedding Approach”. In: *The Review of Financial Studies*, hhaf012. ISSN: 0893-9454.
- Jha, M. et al. (Oct. 4, 2024). *Harnessing Generative AI for Economic Insights*. <https://ssrn.com/abstract=4976759> or <http://dx.doi.org/10.2139/ssrn.4976759>. Available at SSRN.
- (Feb. 2025b). “ChatGPT and Corporate Policies”. Chicago Booth Research Paper No. 23-15, Fama-Miller Working Paper, University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2023-103.
- Jia, N. et al. (2024). “When and How Artificial Intelligence Augments Employee Creativity”. In: *Academy of Management Journal* 67.1, pp. 5–32.
- Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.
- Kaplan, J. et al. (2020). *Scaling Laws for Neural Language Models*.
- Kok, T. de (2025). “ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research”. In: *Management Science* 0.0.
- Korinek, A. (Jan. 2023). “Generative AI for Economic Research: Use Cases and Implications for Economists”. In: *Journal of Economic Literature* 61.4, pp. 1281–1317.
- Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings”. In: *American Sociological Review* 84.5, pp. 905–949. ISSN: 00031224, 19398271.
- Le, Q. V. and T. Mikolov (2014). *Distributed Representations of Sentences and Documents*.
- Li, N. et al. (2024). *EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities*.

- Lippert, S. et al. (2024). “Can large language models help predict results from a complex behavioural science study?” In: *Royal Society Open Science* 11.9.
- Lopez-Lira, A. and Y. Tang (Apr. 2023). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. Available at SSRN. URL: <https://ssrn.com/abstract=4412788>.
- Ludwig, J. and S. Mullainathan (2024). “Machine learning as a tool for hypothesis generation”. In: *The Quarterly Journal of Economics* 139.2, pp. 751–827.
- Ludwig, J., S. Mullainathan, and A. Rambachan (Jan. 2025). *Large Language Models: An Applied Econometric Framework*. Working Paper 33344. National Bureau of Economic Research.
- Luo, X., A. Rechardt, G. Sun, et al. (Feb. 2025). “Large language models surpass human experts in predicting neuroscience results”. In: *Nature Human Behaviour* 9. Received: 19 March 2024, Accepted: 02 October 2024, Published: 27 November 2024, pp. 305–315.
- Manning, B. S., K. Zhu, and J. J. Horton (Apr. 2024). *Automated Social Science: Language Models as Scientist and Subjects*. Working Paper 32381. National Bureau of Economic Research.
- Manzoor, E. et al. (2024). *Workshop on AI and Behavioral Science (AIBS 2024)*. URL: <https://ai-behavioral-science.github.io/2024> (visited on 04/29/2025).
- McCarthy, J. et al. (Dec. 2006). “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955”. In: *AI Magazine* 27.4, p. 12.
- Mei, Q. et al. (2024). “A Turing test of whether AI chatbots are behaviorally similar to humans”. In: *Proceedings of the National Academy of Sciences* 121.9, e2313925121.
- Meng, J. (2024). “AI emerges as the frontier in behavioral science”. In: *Proceedings of the National Academy of Sciences* 121.10, e2401336121.
- Meng, K. et al. (2023). *Locating and Editing Factual Associations in GPT*.
- Mikolov, T. et al. (2013). *Efficient Estimation of Word Representations in Vector Space*.
- Mills, C. W. (2000). *The Sociological Imagination*. Oxford: Oxford University Press.
- Motoki, F., V. Pinho Neto, and V. Rodrigues (2024). “More human than human: measuring ChatGPT political bias”. In: *Public Choice* 198.1, pp. 3–23. ISSN: 1573-7101.
- Mullainathan, S. and J. Spiess (2017). “Machine Learning: An Applied Econometric Approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.
- Noy, S. and W. Zhang (2023). “Experimental evidence on the productivity effects of generative artificial intelligence”. In: *Science* 381.6654, pp. 187–192.
- Ouyang, L. et al. (2022). *Training language models to follow instructions with human feedback*.
- Qin, X., M. Huang, and J. Ding (June 2024). *AITurk: Using ChatGPT for Social Science Research*. Available at SSRN. URL: <https://ssrn.com/abstract=4922861>.

- Qu, Y. and J. Wang (Aug. 28, 2024). “Performance and biases of Large Language Models in public opinion simulation”. In: *Humanities and Social Sciences Communications* 11.1, p. 1095. ISSN: 2662-9992.
- Raffel, C. et al. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*.
- Reimers, N. and I. Gurevych (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.
- Shiller, R. J. (Apr. 2017). “Narrative Economics”. In: *American Economic Review* 107.4, pp. 967–1004.
- Siano, F. (2025). “The news in earnings announcement disclosures: Capturing word context using LLM methods”. In: *Management Science*.
- Sutton, R. S. (Mar. 2019). *The Bitter Lesson*. Blog post.
- Tranchero, M. et al. (Oct. 2024). *Theorizing with Large Language Models*. Working Paper 33033. National Bureau of Economic Research.
- Vaccaro, M., A. Almaatouq, and T. Malone (2024). “When combinations of humans and AI are useful: A systematic review and meta-analysis”. In: *Nature Human Behaviour*, pp. 1–11.
- Vaswani, A. et al. (2023). *Attention Is All You Need*.
- Wager, S. and S. Athey (2018). “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests”. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.
- Wang, M. et al. (2024). *Knowledge Mechanisms in Large Language Models: A Survey and Perspective*.
- Wang, S. et al. (2024). *Knowledge Editing for Large Language Models: A Survey*.
- Wei, J. et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 24824–24837.
- Zhao, Y. et al. (2024). *Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions*.